MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0024
Measure Title: Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents (WCC)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: Percentage of patients 3-17 years of age who had an outpatient visit with a primary care physician (PCP) or an OB/GYN and who had evidence of the following during the measurement year:
- Body mass index (BMI) percentile documentation
- Counseling for nutrition
- Counseling for physical activity

Developer Rationale: Obesity and poor nutrition or physical activity habits in children and adolescents are associated both with immediate health concerns and longer-term morbidity, e.g., asthma, orthopedic problems, adverse cardiovascular and metabolic outcomes, and mental health issues. For children who are overweight or obese, obesity in adulthood is likely to be more severe and lead to obesity-related morbidity, i.e. type 2 diabetes.

Numerator Statement: Patients who had evidence of the following during the measurement year: a body mass index (BMI) percentile documentation, counseling for nutrition, counseling for physical activity.
Denominator Statement: Patients 3-17 years of age with at least one outpatient visit with a primary care physician (PCP) or OB-GYN during the measurement year.
Denominator Exclusions: The measure excludes female patients who have a diagnosis of pregnancy and patients who use hospice services during the measurement year.

Measure Type: Process
Data Source: Claims, Electronic Health Records, Paper Medical Records
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Oct 19, 2012

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- This measure focuses on the patients ages 3-17 years with at least one outpatient visit with a primary care physician (PCP) or OB-GYN who received a body mass index (BMI) percentile documentation, counseling for nutrition, and counseling for physical activity during the measurement year.
- The developer provides the following logic model to support the measure: Children and adolescents have an outpatient visit with a primary care provider (PCP) or obstetrician/gynecologist (OB/GYN) → Body mass index (BMI) percentile documentation, counseling for nutrition, and counseling for physical activity occur → Obesity in children and adolescents is 1) prevented or 2) identified and addressed → Morbidity associated with obesity is prevented → Health outcomes are improved
- The developer cites a United States Preventative Services Task Force (USPSTF) recommendation that clinicians screen for obesity in children and adolescents 6 years and older and offer or refer them to comprehensive, intensive behavioral interventions to promote improvements in weight status. The recommendation received a B grade, which means that USPSTF concludes with moderate certainty that the net benefit of screening for obesity in children and adolescents 6 years and older and offering or referring them to comprehensive, intensive behavioral interventions to promote improvements in weight status is moderate.
- The systematic review that supports the measure includes 140 randomized control trials (RCTs) of good or fair quality related to various aspects of the effectiveness of weight loss and weight management interventions.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer updated the Evidence form to provide the 2017 USPSTF guidelines (from the 2010 recommendation). Both the 2010 and 2017 guidelines recommend that clinicians screen for obesity in children and adolescents 6 years and older and offer or refer them to comprehensive, intensive behavioral interventions to promote improvements in weight status. Both guidelines received an B rating, meaning that the USPSTF concludes with moderate certainty that the net benefit of the measure focus is moderate.

Questions for the Committee:

- *The developer provided an updated 2017 USPSTF guideline to support the measure focus. Does the Committee agree that the measure reflects the current USPSTF recommendation? Does the Committee wish to discuss why the measure is specified for a different age group than stated in the guideline?*

Guidance from the Evidence Algorithm

Measure a health outcome (Box 1) No → Assess performance of intermediate outcome, process, or structure(Box 3) Yes → Summary of QQC provided (Box 4) Yes → moderate certainty that the net benefit is substantial (Box 5) → Moderate rating

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer provides the data below, which are extracted from HEDIS and reflect the most recent years of performance for this measure.
- From 2014 to 2016, performance rates for this measure have shown slight improvement across commercial and Medicaid plans.

BMI Percentile Documentation Mean			
Measurement Year	**2014**	**2015**	**2016**
Commercial- Ages 3-11	51.2%	55.1%	59.7%
Commercial- Ages 12-17	51.5%	52.0%	56.8%
Commercial- Total	51.3%	53.7%	58.4%
Medicaid- Ages 3-11	63.6%	64.8%	69.8%
Medicaid- Ages 12-17	64.7%	63.2%	67.9%
Medicaid- Total	64.0%	64.4%	69.1%

Counseling for Nutrition Mean			
Measurement Year	**2014**	**2015**	**2016**
Commercial- Ages 3-11	53.2%	55.4%	58.0%
Commercial- Ages 12-17	46.8%	49.4%	51.8%
Commercial- Total	50.5%	52.8%	55.3%
Medicaid- Ages 3-11	62.2%	61.6%	66.5%
Medicaid- Ages 12-17	57.5%	57.3%	63.2%
Medicaid- Total	60.5%	60.2%	65.3%

Counseling for Physical Activity Mean			
Measurement Year	**2014**	**2015**	**2016**
Commercial- Ages 3-11	46.9%	47.6%	48.7%
Commercial- Ages 12-17	48.8%	50.4%	52.4%
Commercial- Total	47.7%	48.7%	50.2%
Medicaid- Ages 3-11	52.6%	52.4%	56.0%
Medicaid- Ages 12-17	55.2%	55.2%	61.0%
Medicaid- Total	53.5%	53.4%	57.6%

Disparities

- HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status if the data are available to a plan.
- The developer provided the following disparities information from the literature:
    - The prevalence of obesity is about 21-25% among African American and Hispanic children 6 years and older, compared to 3.7% among Asian girls aged 6 to 11 years, and 20.9% among non-Hispanic white adolescent girls. (O'Connor et al, 2017; Ogden et al, 2012)
    - Studies have found the percentage of obese/overweight children and adolescents to be greater in communities with lower household incomes (Eagle et al, 2012)
    - Studies also have found geographic disparities in the prevalence of obesity. Obesity rates are higher among rural children than urban children (the odds of obesity are 26% greater in rural children compared to their urban

counterparts). Rural adolescents are also more likely to be obese and eat fewer fruits and vegetables than urban adolescents. (Gustafson, 2017; Johnson and Johnson, 2015)

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**
Evaluators: NQF Staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Staff analysis of Scientific Acceptability

Questions for the Committee regarding reliability:

- *This is a maintenance measure and staff is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *This is a maintenance measure and staff is satisfied with the validity testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Data are generated or collected and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
- To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records, and registry). The developer anticipates that as electronic health records become more widespread the reliance on paper record review will decrease.

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- This measure is used in the Quality Payment Program (QPP) and is included in the core set of health quality measures for children enrolled in Medicaid/Children's Health Insurance Program (CHIP), to be reported at the state level.
- The measure also is used in several ratings and benchmarking programs, including the NCQA State of Health Care annual report, NCQA health plan ratings/report cards, NCQA Quality Compass, and the Qualified Health Plan (QHP) Quality rating System.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

Additional Feedback:

- Questions received on the measure have generally centered around clarification on whether certain notations in medical record documentation are sufficient to meet the measure specifications. Other questions have sought clarification about what type of provider needs to conduct the various numerator components. The developer has provided minor clarifications about the measure during the annual update process in order to address questions received through the NCQA Policy Clarification Support system.

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- From 2014 to 2016, performance rates for this measure have shown slight improvement across commercial and Medicaid plans. In 2016, commercial plans on average had performance rates of 58%, 55% and 50%% for BMIpercentile documentation, nutrition counseling and physical activity counseling, respectively. In 2016, Medicaid plans on average had performance rates of 69%, 65% and 58% for BMI percentile documentation, nutrition counseling and physical activity counseling, respectively.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

Potential harms

- The developer reported that no unexpected findings were identified during testing or since implementation of this measure.

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- N/A

Harmonization

- N/A

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Scientific Acceptability

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number:
Measure Title:

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?
   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?
   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
   ☒Yes (go to Question #5)
   ☐No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☒Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
   Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☒High (go to Question #8)
   ☐Moderate (go to Question #8)
   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☐Yes (go to Question #9)
   ☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☐Yes (go to Question #10)

☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

    ☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?
   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.
   ☐Yes (go to Question #2)
   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?
   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?
   ☐Yes (please explain below then go to Question #3)
   ☐No (go to Question #3)
   ☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☐Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?
   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*
   ☐Yes (please explain below then go to Question #4)
   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
   ☐Yes (please explain below then go to Question #5)
   ☐No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
   ☐Yes (please explain below then go to Question #6)
   ☐No (go to Question #6)

☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☐No (go to Question #7)

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☒Yes (go to Question #11)
    ☐No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☒Yes (go to Question #12)
    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #14)
☒Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

13. Was other validity testing reported?

☐Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

TIPS: Prior validity studies of the same data elements may be submitted

☐Yes (go to Question #15)
☒No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #16)
☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0024

Measure Title: Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 11/15/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the

step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)

☐ Intermediate clinical outcome (*e.g., lab value*):

☒ Process: Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Children and adolescents have an outpatient visit with a primary care provider (PCP) or obstetrician/gynecologist (OB/GYN) >> Body mass index (BMI) percentile documentation, counseling for nutrition, and counseling for physical activity occur >> Obesity in children and adolescents is 1) prevented or 2) identified and addressed >> Morbidity associated with obesity is prevented >> Health outcomes are improved

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

N/A

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

N/A

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☒ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	US Preventive Services Task Force (USPSTF). Screening for obesity in children and adolescents. US Preventive Services Task Force Recommendation Statement. *JAMA*. 2017;317(23):2417-2426. https://jamanetwork.com/journals/jama/fullarticle/2632511
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	"The USPSTF recommends that clinicians screen for obesity in children and adolescents 6 years and older and offer or refer them to comprehensive, intensive behavioral interventions to promote improvements in weight status. (B recommendation)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	"The USPSTF concludes with moderate certainty that the net benefit of screening for obesity in children and adolescents 6 years and older and offering or referring them to comprehensive, intensive behavioral interventions to promote improvements in weight status is moderate." The USPSTF included studies that were fair- or good-quality studies. The following text is directly quoted from the USPSTF eTable1. Quality Assessment Criteria

Study Design	Adapted Quality Criteria
Randomized and non-randomized controlled-trials, adapted from the U.S Preventive Services Task Force methods (Harris et al, 2001)	• Valid random assignment? • Was allocation concealed? • Was eligibility criteria specified? • Were groups similar at baseline? • Was there a difference in attrition between groups? • Were outcome assessors blinded? • Were measurements equal, valid and reliable? • Was there intervention fidelity? • Was there risk of contamination? • Was there adequate adherence to the intervention? • Were the statistical methods acceptable? • Was the handling of missing data appropriate? • Was there acceptable follow-up? • Was there evidence of selective reporting of outcomes?

- Good quality studies generally meet all quality criteria.
- Fair quality studies do not meet all the criteria but do not have critical limitations that could invalidate study findings.

Critical appraisal of studies using *a priori* quality criteria are conducted independently by at least two reviewers. Disagreements in final quality assessment are resolved by consensus, and, if needed, consultation with a third independent reviewer.

	Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med*. 2001;20(3 Suppl):21-35.
Provide all other grades and definitions from the evidence grading system	• Poor quality studies have a single fatal flaw or multiple important limitations that could invalidate study findings.
Grade assigned to the **recommendation** with definition of the grade	Grade: B "The USPSTF recommends this service. There is high certainty that the net benefit is moderate, or there is moderate certainty that the net benefit is moderate to substantial."
Provide all other grades and definitions from the recommendation grading system	A. The USPSTF recommends this service. There is high certainty that the net benefit is substantial. C. The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small. D: The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. I: The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	QUANTITY Key Question 1: Do screening programs for obesity in children and adolescents lead to reductions in excess weight or age-associated excess weight gain, improve health outcomes during childhood, or reduce incidence of obesity in adulthood? • No identified studies meeting the inclusion criteria addressed this key question. Key Question 2: Does screening for obesity in children and adolescents have adverse effects? • No identified studies meeting the inclusion criteria addressed this key question. Key Question 3: Do lifestyle-based weight loss interventions for children and adolescents embedded in

	primary care, or to which primary care physicians refer, improve health outcomes during childhood or reduce incidence of obesity in adulthood? • 10 RCTs of lifestyle-based weight loss reported measures of health-related quality of life, functioning or both using the Pediatric Quality of Life Inventory, the Child Health Questionnaire or DISABKIDS. • 1 RCT of lifestyle-based weight loss reported changes in physical functioning with a larger effect size. Key Question 4: Do [lifestyle-based weight loss] interventions for children and adolescents that are embedded in primary care, or to which primary care physicians refer, reduce excess weight or age-associated excess weight gain? "Lifestyle-based weight loss interventions provided at least dietary counseling and some information about behavior change principles, and most provided information related to physical activity or sedentary behavior." • 39 RCTs • 3 CCTs Key Question 4a: Do [lifestyle-based] weight management interventions affect cardiometabolic measures? • 6 reporting measures of blood pressure • 4 reporting measures of lipids • 4 reporting measures of fasting plasma glucose Key Question 4b: Are there common components of efficacious interventions? • Due to the limited number of studies, variation in reported outcomes and similar effect sizes across studies, there was insufficient data to address this key question. Key Question 4c: Does efficacy differ by key patient subgroups (i.e., age, race/ethnicity, sex, degree of excess weight, and socioeconomic status)? • Due to the limited number of studies, variation in reported outcomes and similar effect sizes across studies, there was insufficient data to address this key question.

	Key Question 5: Do weight management interventions for children and adolescents have adverse effects? • 5 RCTs reporting any adverse events • 5 RCTs reporting measures of disordered eating or body dissatisfaction QUALITY The data for this report was extracted from fair- and good-quality trials. Key Question 3: Do lifestyle-based weight loss interventions for children and adolescents embedded in primary care, or to which primary care physicians refer, improve health outcomes during childhood or reduce incidence of obesity in adulthood? • 5 RCTs of good quality • 6 RCTs of fair quality Key Question 4: Do [lifestyle-based weight loss] interventions for children and adolescents that are embedded in primary care, or to which primary care physicians refer, reduces excess weight or age-associated excess weight gain? Lifestyle-based weight loss interventions provided at least dietary counseling and some information about behavior change principles, and most provided information related to physical activity or sedentary behavior." • 8 RCTs of good quality • 34 trials of fair quality Key Question 4a: Do [lifestyle-based] weight management interventions affect cardiometabolic measures? • The evidence review did not report the quality for studies addressing this question. Key Question 5: Do weight management interventions for children and adolescents have adverse effects? • 4 RCTs of good quality • 6 RCTs of fair quality
Estimates of benefit and consistency across studies	The following text is quoted directly from the USPSTF recommendation statement by O'Connor et al, 2017. "There was no direct evidence on the benefits or harms of screening children and adolescents for excess weight, but

	a fairly large and recent body of evidence suggests that lifestyle-based weight loss programs with at least 26 hours of contact are likely to promote reductions in excess weight in children and adolescents. The literature also revealed no evidence of these programs causing harm. Relative reductions in BMI *z* score of 0.20 or more were typical, but the absolute amount of weight loss was highly variable within studies, suggesting a wide possible range of benefit. Those with the most contact hours also demonstrated approximately 6–mm Hg reductions in SBP [systolic blood pressure] relative to the control groups, smaller reductions in DBP [diastolic blood pressure], and some improvement in insulin and glucose measures, but typically no improvements in levels of fasting plasma glucose or lipids. Behavior-based interventions with fewer estimated hours of contact rarely demonstrated benefit, although limited evidence suggested that briefer interventions may be effective in children who are overweight but who do not have obesity. Estimated hours of contact was the only characteristic clearly related to effect size, with larger effects seen in trials with more contact hours."
What harms were identified?	The following text is quoted directly from the USPSTF recommendation statement by O'Connor et al, 2017. "There was no direct evidence on the benefits or harms of screening children and adolescents for excess weight, but a fairly large and recent body of evidence suggests that lifestyle-based weight loss programs with at least 26 hours of contact are likely to promote reductions in excess weight in children and adolescents. The literature also revealed no evidence of these programs causing harm."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	One related study (Shreve et al, 2017) has been published since the publication of this systematic review. The conclusion of this study does not contradict the conclusion from the systematic review. Shreve M, Scott A, Vowell Johnson K. Adequately addressing pediatric obesity: challenges faced by primary care providers. *Southern Medical Journal*. 2017;110(7):486-490.

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

NQF_-_WCC_-_Evidence_Attachment.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Obesity and poor nutrition or physical activity habits in children and adolescents are associated both with immediate health concerns and longer-term morbidity, e.g., asthma, orthopedic problems, adverse cardiovascular and metabolic outcomes, and mental health issues. For children who are overweight or obese, obesity in adulthood is likely to be more severe and lead to obesity-related morbidity, i.e. type 2 diabetes.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are presented at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data are shown by year and product line (i.e. commercial, Medicaid).

Commercial – BMI Percentile – Ages 3-11 Years

YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR

2014 | 51.20% | 27.1% | 0.0% | 99.5% | 3.6% | 38.2% | 55.7% | 70.3% | 84.5% | 32.1

2015 | 55.1% | 24.7% | 0.4% | 100.0% | 9.5% | 43.7% | 59.0% | 72.9% | 83.9% | 29.2

2016 | 59.7% | 24.2% | 0.7% | 100.0% | 15.2% | 50.3% | 64.3% | 76.2% | 86.8% | 25.9

Commercial – BMI Percentile – Ages 12-17 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 51.5% | 26.8% | 0.0% | 100.0% | 3.9% | 41.1% | 56.9% | 69.5% | 81.8% | 28.4
2015 | 52.0% | 24.3% | 0.3% | 100.0% | 9.3% | 40.5% | 55.0% | 68.0% | 81.0% | 27.5
2016 | 56.8% | 23.7% | 1.2% | 100.0% | 11.8% | 45.6% | 59.9% | 73.8% | 84.2% | 28.2

Commercial – BMI Percentile - Total
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 51.3% | 26.9% | 0.0% | 99.2% | 3.6% | 40.1% | 56.2% | 70.2% | 83.0% | 30.1
2015 | 53.7% | 24.4 % | 0.5% | 99.1% | 9.1% | 42.3% | 57.3% | 71.2% | 82.2% | 28.9
2016 | 58.4% | 23.8% | 1.1% | 100.0% | 14.1% | 47.9% | 62.4% | 74.9% | 85.2% | 27

Commercial – Counseling for Nutrition – Ages 3-11 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 53.2% | 26.8% | 0.0% | 98.6% | 3.6% | 46.9% | 59.8% | 70.1% | 81.3% | 23.2
2015 | 55.4% | 23.9% | 0.4% | 98.4% | 6.5% | 47.2% | 60.3% | 71.0% | 81.2% | 23.8
2016 | 58.0% | 24.0% | 0.3% | 100.0% | 8.9% | 50.6% | 63.3% | 73.8% | 83.4% | 23.2

Commercial – Counseling for Nutrition – Ages 12-17 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 46.8% | 25.3% | 0.0% | 98.9% | 2.9% | 36.0% | 51.6% | 62.4% | 73.7% | 26.4
2015 | 49.4% | 22.8% | 0.1% | 100.0% | 7.7% | 37.8% | 53.2% | 64.1% | 75.4% | 26.3
2016 | 51.8% | 23.0% | 0.4% | 100.0% | 7.2% | 42.2% | 54.6% | 66.0% | 77.8% | 23.8

Commercial – Counseling for Nutrition - Total
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 50.5% | 26.0% | 0.0% | 98.3% | 3.2% | 41.8% | 56.8% | 67.1% | 77.9% | 25.3
2015 | 52.8% | 23.3% | 0.3% | 99.1% | 6.0% | 43.6% | 57.6% | 67.9% | 79.2% | 24.3
2016 | 55.3% | 23.3% | 0.4% | 100% | 8.5% | 46.2% | 59.7% | 70.3% | 79.7% | 24.1

Commercial – Counseling for Physical Activity – Ages 3-11 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 46.9% | 24.8% | 0.0% | 98.6% | 2.9% | 37.2% | 51.9% | 63.7% | 74.2% | 26.5
2015 | 47.6% | 22.3% | 0.0% | 98.4% | 4.8% | 37.2% | 50.9% | 62.7% | 72.7% | 25.5
2016 | 48.7% | 23.1% | 0.0% | 100.0% | 1.3% | 38.8% | 52.6% | 63.6% | 75.1% | 24.8

Commercial – Counseling for Physical Activity – Ages 12-17 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 48.8% | 25.8% | 0.0% | 100.0% | 2.1% | 38.3% | 54.5% | 65.7% | 76.9% | 27.4
2015 | 50.4% | 23.0% | 0.0% | 100.0% | 5.3% | 40.6% | 54.4% | 65.4% | 75.9% | 24.8
2016 | 52.4% | 23.0% | 0.0% | 100.0% | 5.2% | 43.9% | 55.7% | 67.2% | 78.0% | 23.3

Commercial – Counseling for Physical Activity - Total
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 47.7% | 25.0% | 0.0% | 98.3% | 2.4% | 38.7% | 53.1% | 64.6% | 73.2% | 25.9
2015 | 48.7% | 22.5% | 0.0% | 99.1% | 5.3% | 38.7% | 52.4% | 63.1% | 74.0% | 24.4
2016 | 50.2% | 23.0% | 0.0% | 100% | 2.8% | 41.0% | 53.8% | 64.4% | 77.1% | 23.4

Medicaid – BMI Percentile – Ages 3-11 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 63.6% | 19.1% | 2.2% | 99.6% | 36.8% | 50.7% | 66.9% | 77.5% | 86.3% | 26.8
2015 | 64.8% | 18.6% | 1.7% | 99.4% | 41.3% | 55.0% | 68.2% | 78.4% | 86.3% | 23.4
2016 | 69.8% | 16.6% | 6.6% | 100.0% | 51.2% | 61.1% | 72.4% | 80.9% | 87.8% | 19.8

Medicaid – BMI Percentile – Ages 12-17 Years

YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 64.7% | 18.3% | 3.7% | 100.0% | 40.0% | 52.1% | 67.5% | 79.5% | 86.4% | 27.4
2015 | 63.2% | 18.9% | 1.6% | 100.0% | 39.2% | 51.6% | 65.7% | 76.8% | 85.2% | 25.2
2016 | 67.9% | 16.7% | 8.2% | 100.0% | 47.4% | 58.9% | 70.5% | 79.5% | 85.8% | 20.6

Medicaid – BMI Percentile - Total
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 64.0% | 18.6% | 2.6% | 99.6% | 38.9% | 51.3% | 67.2% | 78.0% | 85.6% | 26.7
2015 | 64.4% | 18.5% | 1.7% | 99.4% | 40.1% | 54.5% | 67.5% | 77.8% | 86.4% | 23.3
2016 | 69.1% | 16.6% | 7.0% | 100% | 48.9% | 60.2% | 72.2% | 80.5% | 87.5% | 20.3

Medicaid – Counseling for Nutrition – Ages 3-11 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 62.2% | 17.7% | 0.4% | 98.8% | 43.3% | 54.3% | 63.0% | 73.8% | 80.3% | 19.5
2015 | 61.6% | 17.5% | 0.4% | 98.1% | 43.5% | 53.0% | 63.3% | 73.4% | 80.2% | 20.4
2016 | 66.5% | 17.2% | 0.3% | 100.0% | 50.4% | 58.8% | 68.9% | 77.8% | 83.9% | 19

Medicaid – Counseling for Nutrition – Ages 12-17 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 57.5% | 18.6% | 0.8% | 100.0% | 36.8% | 47.8% | 58.3% | 71.5% | 77.8% | 23.7
2015 | 57.3% | 17.5% | 0.8% | 97.7% | 40.1% | 47.8% | 57.4% | 68.4% | 78.7% | 20.6
2016 | 63.2% | 17.4% | 0.6% | 100.0% | 44.5% | 55.7% | 65.0% | 74.2% | 81.5% | 18.5

Medicaid – Counseling for Nutrition - Total
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 60.5% | 17.8% | 0.5% | 98.1% | 41.4% | 52.0% | 61.4% | 72.9% | 79.6% | 20.9
2015 | 60.2% | 17.2% | 0.5% | 97.6% | 42.9% | 51.8% | 62.6% | 70.9% | 79.5% | 19.1
2016 | 65.3% | 17.2% | 0.5% | 98.5% | 48.6% | 58.6% | 68.0% | 76.6% | 82.5% | 18

Medicaid – Counseling for Physical Activity – Ages 3-11 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 52.6% | 17.4% | 0.0% | 98.2% | 34.8% | 42.9% | 53.4% | 63.9% | 71.8% | 21
2015 | 52.4% | 17.0% | 0.0% | 98.1% | 35.6% | 43.6% | 54.0% | 62.2% | 71.3% | 18.6
2016 | 56.0% | 17.7% | 0.0% | 100.0% | 39.4% | 47.1% | 57.2% | 66.6% | 76.1% | 19.5

Medicaid – Counseling for Physical Activity – Ages 12-17 Years
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 55.2% | 18.1% | 0.0% | 100.0% | 35.7% | 46.5% | 56.3% | 66.2% | 75.4% | 19.7
2015 | 55.2% | 17.5% | 0.1% | 97.2% | 37.0% | 46.5% | 55.8% | 65.4% | 74.6% | 18.9
2016 | 61.0% | 16.6% | 0.5% | 100.0% | 45.1% | 54.0% | 62.1% | 70.7% | 78.3% | 16.7

Medicaid – Counseling for Physical Activity - Total
YEAR | MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 53.5% | 17.3% | 0.0% | 98.1% | 35.8% | 44.2% | 53.9% | 64.4% | 71.5% | 20.2
2015 | 53.4% | 16.8% | 0.0% | 97.6% | 35.9% | 45.1% | 55.4% | 63.5% | 71.6% | 18.4
2016 | 57.6% | 17.1% | 0.4% | 100% | 41.6% | 49.1% | 59.3% | 67.6% | 75.4% | 18.5

In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 47.0 million Medicaid beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans that reported the measure and the median eligible population for the measure across health plans.

Commercial – BMI Percentile - Total
YEAR | N Plans | Median Denominator Size per plan
2014 | 381 | 411
2015 | 409 | 411
2016 | 406 | 411

Commercial – Counseling for Nutrition - Total
YEAR | N Plans | Median Denominator Size per plan
2014 | 379 | 411
2015 | 406 | 411
2016 | 402 | 411

Commercial – Counseling for Physical Activity - Total
YEAR | N Plans | Median Denominator Size per plan
2014 | 376 | 411
2015 | 404 | 411
2016 | 396 | 411

Medicaid – BMI Percentile - Total
YEAR | N Plans | Median Denominator Size per plan
2014 | 208 | 411
2015 | 244 | 411
2016 | 219 | 411

Medicaid – Counseling for Nutrition - Total
YEAR | N Plans | Median Denominator Size per plan
2014 | 208 | 411
2015 | 244 | 411
2016 | 246 | 411

Medicaid – Counseling for Physical Activity - Total
YEAR | N Plans | Median Denominator Size per plan
2014 | 208 | 411
2015 | 244 | 411
2016 | 246 | 411

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4**
Recognizing disparities in obesity prevalence, nutrition and physical activity behaviors can help ensure successful interventions to curb childhood obesity and prevent morbidity. Some behaviors are highly predictive of obesity, e.g. low levels of moderate

physical activity and poor dietary intake. Although overall obesity rates in children and adolescents have stabilized over the last decade, obesity rates continue to increase in certain populations, i.e. African American girls and Hispanic boys. The prevalence of obesity is about 21 percent to 25 percent among African American and Hispanic children 6 years and older, compared to 3.7 percent among Asian girls aged 6 to 11 years, and 20.9 percent among non-Hispanic white adolescent girls. (O'Connor et al, 2017; Ogden et al, 2012) Studies have found the percentage of obese/overweight children and adolescents to be greater in communities with lower household incomes. Children living in lower income communities exhibit poorer dietary and physical activity behaviors, i.e. increased fried food consumption and increased TV/video time (in Michigan sixth graders, frequency of fried food consumed doubles from 0.23 to 0.54 as household income decreases, and TV/video time triples from 0.55 to 2.00 hours daily as household income decreases). (Eagle et al, 2012) Studies have also found geographic disparities in the prevalence of obesity. Obesity rates are higher among rural children than urban children (the odds of obesity are 26 percent greater in rural children compared to their urban counterparts). Rural adolescents are also more likely to be obese and eat fewer fruits and vegetables than urban adolescents. (Gustafson, 2017; Johnson and Johnson, 2015)

Eagle TF, Sheetz A, Gurm R, et al. Understanding childhood obesity in America: linkages between household income, community resources, and children's behaviors. American Heart Journal. 2012;163(5):836-843.

Gustafson A, Pitts SJ, McDonald J, et al. Direct effects of the home, school, and consumer food environments on the association between food purchasing patterns and dietary intake among rural adolescents in Kentucky and North Carolina. International Journal of Environmental Research and Public Health. 2017;14(10)1255.

Johnson JA and Johnson AM. Urban-rural differences in childhood and adolescent obesity in the United States: a systematic review and meta-analysis. Child Obesity. 2015;11(3)233-41.

O'Connor EA, Evans CV, Burda BU, Walsh ES, Eder M, Lozano P. Screening for Obesity and Intervention for Weight Management in Children and Adolescents: A Systematic Evidence Review for the US Preventive Services Task Force. Evidence Synthesis No. 150. Rockville, MD: Agency for Healthcare Research and Quality; 2017. AHRQ publication 15-05219-EF-1.

Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011-2012. JAMA. 2014;311(8):806-814.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Non-Condition Specific***(check all the areas that apply):*
Primary Prevention

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
N/A

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0024_WCC_Value_Sets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
No changes

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients who had evidence of the following during the measurement year: a body mass index (BMI) percentile documentation, counseling for nutrition, counseling for physical activity.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
ADMINISTRATIVE:
BMI Percentile: Patients with a BMI percentile* (BMI Percentile Value Set) during the measurement year

*Because BMI norms for youth vary with age and gender, this measure evaluates whether BMI percentile is assessed rather than an absolute BMI value

Counseling for Nutrition: Patients with counseling for nutrition (Nutrition Counseling Value Set) during the measurement year

Counseling for Physical Activity: Patients with counseling for physical activity (Physical Activity Counseling Value Set) during the measurement year

---
MEDICAL RECORD:
BMI Percentile:
Patients with documentation in the medical record of a BMI percentile during the measurement year. Documentation must include height, weight and BMI percentile during the measurement year. The height, weight and BMI percentile must be from the same data source. Either of the following meets criteria for BMI percentile:
• BMI percentile documented as a value (e.g., 85th percentile).

• BMI percentile plotted on an age-growth chart.

The percentile ranking based on the CDC's BMI-for-age growth charts, which indicates the relative position of the patient's BMI number among others of the same gender and age.

Only evidence of the BMI percentile or BMI percentile on an age-growth chart meets criteria.

Ranges and thresholds do not meet criteria for this indicator. A distinct BMI percentile is required for numerator compliance. Documentation of >99% or <1% meet criteria because a distinct BMI percentile is evident (i.e., 100% or 0%).

Counseling for Nutrition:
Patients with documentation in the medical record of counseling for nutrition or referral for nutrition education during the measurement year. Documentation must include a note indicating the date and at least one of the following:
• Discussion of current nutrition behaviors (e.g., eating habits, dieting behaviors).
• Checklist indicating nutrition was addressed.
• Counseling or referral for nutrition education.
• Patient received educational materials on nutrition during a face-to-face visit.
• Anticipatory guidance for nutrition.
• Weight or obesity counseling.

Counseling for Physical Activity:
Patients with documentation in the medical record of counseling for physical activity or referral for physical activity during the measurement year. Documentation must include a note indicating the date and at least one of the following:
• Discussion of current physical activity behaviors (e.g., exercise routine, participation in sports activities, exam for sports participation).
• Checklist indicating physical activity was addressed.
• Counseling or referral for physical activity.
• Patient received educational materials on physical activity during face-to-face visit.
• Anticipatory guidance specific to the child's physical activity.
• Weight or obesity counseling.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 3-17 years of age with at least one outpatient visit with a primary care physician (PCP) or OB-GYN during the measurement year.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients 3-17 years of age as of December 31 of the measurement year with an outpatient visit (Outpatient Value Set) with a PCP or an OB/GYN during the measurement year.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
The measure excludes female patients who have a diagnosis of pregnancy and patients who use hospice services during the measurement year.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
Exclude female patients who have a diagnosis of pregnancy (Pregnancy Value Set) during the measurement year.

Exclude patients who use hospice services any time during the measurement year (Hospice Value Set).

The denominator for all rates must be the same. An organization that excludes these patients must do so for all rates.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
The total population is stratified by age: 3-11 and 12-17 years of age.

Report two age stratifications and a total rate for each of the three indicators.

The total is the sum of the age stratifications.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Step 1. Determine the eligible population. To do so, identify all patients 3-17 years of age who had an outpatient visit (Outpatient Value Set) with a PCP or OB/GYN during the measurement year.
Step 2: Exclude patients with pregnancy diagnosis (Pregnancy Value Set) or who used hospice services (Hospice Value Set) from the eligible population.
Step 3: Determine numerator events. To do so, identify the number of patients in the eligible population who had evidence of BMI percentile documentation (BMI Percentile Value Set), counseling for nutrition (Nutrition Counseling Value Set), and counseling for physical activity (Physical Activity Counseling Value Set) during the measurement year.
Step 4. Calculate the three rates.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Records, Paper Medical Records

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2. Validity – See attached Measure Testing Submission Form
NQF_-_WCC_-_Testing_Attachment.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 0024
Measure Title: Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents
Date of Submission: 11/15/2017
Type of Measure:

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

Instructions

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

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed 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]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record

☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

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? 2016

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

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

1.5. How many and which <u>measured entities</u> 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*)
<u>Sample for measure score reliability testing</u>: The measure score reliability was calculated from HEDIS data that included 246 Medicaid health plans and 406 commercial health plans. The sample included all commercial and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

<u>Sample for construct validity testing</u>: Construct validity was calculated from HEDIS data that included 216 Medicaid health plans and 406 commercial health plans. The sample included all commercial and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

1.6. How many and which <u>patients</u> 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*)
<u>Patient sample for measure score reliability testing</u>: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 47.0 million Medicaid beneficiaries. Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicaid). Below is a description of the

sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Rate	Product Type	Number of Plans	Median number of eligible patients per plan
BMI Percentile	Commercial	406	411
Counseling for Nutrition	Commercial	402	411
Counseling for Physical Activity	Commercial	396	411
BMI Percentile	Medicaid	219	411
Counseling for Nutrition	Medicaid	246	411
Counseling for Physical Activity	Medicaid	246	411

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

Rate	Product Type	Number of Plans	Median number of eligible patients per plan
BMI Percentile	Commercial	406	411
Counseling for Nutrition	Commercial	402	411
Counseling for Physical Activity	Commercial	396	411
BMI Percentile	Medicaid	216	411
Counseling for Nutrition	Medicaid	215	411
Counseling for Physical Activity	Medicaid	215	411

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.

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

Validity was demonstrated through construct validity.

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.
We did not analyze performance by social risk factors.

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

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

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

Rate	Commercial	Medicaid
BMI Percentile	0.999	0.993
Counseling for Nutrition	0.999	0.995
Counseling for Physical Activity	0.999	0.996

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?*)
Interpretation of measure score reliability testing: The testing suggests the measure has high reliability.

2b1. VALIDITY TESTING
2b1.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☒ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
Method of testing construct validity: We tested for construct validity by exploring whether two measures were correlated with each other. For this measure, we specifically hypothesized that Weight Assessment and Counseling for Nutrition and Physical Activity will be positively correlated with Adult BMI Assessment (i.e. plans that have high performance on weight assessment and counseling for nutrition and physical activity on the child measure will have high performance on adult BMI assessment). To test this correlation, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

Method of assessing face validity: NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (measurement advisory panels [MAPs] – whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness, and feasibility. 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 Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

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

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

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

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

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review, and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation, information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.
Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the new year's HEDIS Volume 2.

2b1.3. What were the statistical results from validity testing? (e.g., correlation; t-test)
Results of construct validity testing: The results in Table 1a and Table 1b suggest there is a strong, positive relationship between these rates in commercial plans and a moderate, positive relationship in Medicaid plans.

Table 1a. Pearson Correlation Coefficients* between *Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents*: Commercial Plans, 2016

	Adult BMI Assessment
Weight Assessment and Counseling: BMI Percentile	0.85
Weight Assessment and Counseling: Nutrition Counseling	0.81
Weight Assessment and Counseling: Physical Activity Counseling	0.79

*All correlations are significant at p<0.001

Table 1b. Pearson Correlation Coefficients* between *Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents*: Medicaid Plans, 2016

	Adult BMI Assessment
Weight Assessment and Counseling: BMI Percentile Documentation	0.64
Weight Assessment and Counseling: Nutrition Counseling	0.64
Weight Assessment and Counseling: Physical Activity Counseling	0.65

*All correlations are significant at p<0.001

Results of face validity assessment: Input from our multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity.

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
Interpretation of construct validity testing: 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. The measures had moderately-high to high correlations (correlation coefficients ranging from 0.639 to 0.852), which indicate the measure has good construct validity.

Interpretation of systematic assessment of face validity: The measurement advisory panel showed good agreement that the measures as specified will accurately differentiate quality across plans. Our interpretation of these results is that this measure has sufficient face validity.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b3*

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

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

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

☒ No risk adjustment or stratification

☐ Statistical risk model with risk factors

☐ Stratification by risk categories

☐ Other,

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

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

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample

size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p-value of the test statistic is less than 0.05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However, the method can be used for comparison of any two measured entities.

2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

HEDIS 2017 Variation in Performance across Health Plans

Plan Type	Rate	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
	BMI Percentile	3809	58.4	23.8	14.1	47.9	62.4	74.9	85.2	27.0	<0.001
	Counseling for Nutrition	3783	55.3	23.3	8.5	46.2	59.7	70.3	79.7	24.1	<0.001
	Counseling for Physical Activity	3715	50.2	23.0	2.8	41.0	53.8	64.4	77.1	23.4	<0.001
	BMI Percentile	1158	69.1	16.6	48.9	60.2	72.2	80.5	87.5	20.4	<0.001
	Counseling for Nutrition	1336	65.3	17.2	48.6	58.6	68.0	76.6	82.5	18.1	<0.001
	Counseling for Physical Activity	1336	57.6	17.1	41.6	49.1	59.3	67.6	75.4	18.6	<0.001

EP: Eligible Population, the average denominator size across plans for the measure
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e.*, *what do the results mean in terms of statistical and meaningful differences?*)

Across both plan types and rates, the difference between the 25th and 75th percentile is statistically significant. Overall, these results suggest there are meaningful differences in performance.

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

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

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

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

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
This measure is collected with a complete sample.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
This measure is collected with a complete sample.

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
This measure is collected with a complete sample.

3. Feasibility
Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
3a. Byproduct of Care Processes For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order). **3a.1. Data Elements Generated as Byproduct of Care Processes.** Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry) If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records, and registry). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Some users report burden that is typical of chart review measures.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Broad public use and dissemination of these measures are encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting Health Plan Ratings https://reportcards.ncqa.org/#/health-plans/list Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Medicaid Child Core Set https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/child-core-set/index.html Qualified Health Plan (QHP) Quality Rating System (QRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/2018_QRS_and_QHP_Enrollee_Survey_Technical_Guidance_20171004_508.pdf Payment Program Quality Payment Program https://qpp.cms.gov/ Quality Improvement (external benchmarking to organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx

4a1.1 **For each CURRENT use, checked above (update for <u>maintenance of endorsement</u>), provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2012, a total of 455 Medicare Advantage health plans, 404 commercial health plans, and 136 Medicaid health plans across 50 states were included in the rankings.

NCQA STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012, the report included measures on 11.5 Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

MEDICAID/CHIP CHILD CORE SET: These are a core set of health quality measures for children enrolled in Medicaid/Children's Health Insurance Program (CHIP) to be reported at the state level. The data collected from these measures will help CMS to better understand the quality of health care that children enrolled in Medicaid/CHIP receive nationally.

NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

QUALIFIED HEALTH PLAN (QHP) QUALITY RATING SYSTEM (QRS): This measure is used in the Qualified Health Plan (QHP) Quality Rating System, which provides comparable information to consumers about the quality of health care services and QHP enrollee experience offered in the Marketplaces.

QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible clinicians (ECs).

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notation in medical record documentation is sufficient to meet measure criteria. Other questions have sought clarification about what type of provider needs to conduct the various numerator components.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the CMS Medicaid Child Core Set and the Qualified Health Plan Quality Rating System.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
We have provided minor clarifications about the measure during the annual update process in order to address questions received through the Policy Clarification Support system.

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
From 2014 to 2016, performance rates for this measure have shown slight improvement across commercial and Medicaid plans. In 2016, commercial plans on average had performance rates of 58 percent, 55 percent and 50 percent for BMI percentile documentation, nutrition counseling and physical activity counseling, respectively. In 2016, Medicaid plans on average had performance rates of 69 percent, 65 percent and 58 percent for BMI percentile documentation, nutrition counseling and physical activity counseling, respectively. There is wide variation between the 10th and 90th percentiles—especially in commercial plans, suggesting room for improvement. For example, among commercial plans, the 2016 rate of children who had documentation of physical activity counseling ranged from 3 percent for plans in the 10th percentile to 77 percent for plans in the 90th percentile. Across commercial plans, there is a large gap in performance between the 10th and 25th percentiles for all three components of this measure (BMI percentile documentation, nutrition counseling and physical activity counseling). For example, in 2016, the rate of children who received nutrition counseling was 8.5 percent compared to 46.2 percent for commercial plans in the 10th percentile and 25th percentile, respectively. When stratified by age group (3-11 years and 12-17 years), performance trends for both commercial and Medicaid plans remained consistent with trends observed for the total.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
There were no identified unexpected findings during testing or since implementation of this measure.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits for this measure during testing or since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The NCQA Childhood/Adolescent Obesity MAP advised NCQA during measure development. They evaluated the way staff specified measures, assessed the content validity of measures, and reviewed field test results. As you can see from the list, the MAP consisted of a balanced group of experts, including representatives from health plans and specialty societies. Note that, in addition to the MAP, we also vetted these measures with a host of other stakeholders, as is our process. Thus, our measures are the result of consensus from a broad and diverse group of stakeholders, in addition to the MAP.

Joe Anarella, MPH, Assistant Director, Bureau of Quality Management and Outcomes Research New York State Department of Health
Keith Bachman, MD, Clinical Lead--CMI Weight Management Initiative, Kaiser Permanente Care Management Institute, Oakland
Terry Bazzarre, PhD, Senior Program Officer, The Robert Wood Johnson Foundation
Chris Bolling, MD (Co-Chair), Medical Director, Medical Weight Loss Program, Cincinnati Children's Hospital Medical Center
William Dietz, MD, PhD, STOP Obesity Alliance, George Washington Unviersity
Molly Gee, MEd, LD, RD, Project Manager, Look Ahead Diabetes Study, Baylor College of Medicine; Chair, Obesity Steering Committee, American Dietetic Association
Sandra Hassink, MD, FAAP, Director, Weight Management Program Department of Pediatrics, Division of General Pediatrics, American Academy of Pediatrics

Francine Kaufman, MD, Professor of Pediatrics, Keck School of Medicine, University of Southern California; Head of the Center for Diabetes, Endocrinology and Metabolism, Children's Hospital Los Angeles
Jonathan Klein, MD, MPH (Co-Chair) Associate Professor of Pediatrics and of Community and Preventive Medicine, University of Rochester; Director, American Academy of Pediatrics, Julius B. Richmond Center of Excellence
Nancy F. Krebs, MD, Professor of Pediatrics University of Colorado School of Medicine, Medical Director, Department of Coordinated Nutrition Services at the Children's Hospital
Catherine MacLean, MD, PhD, VA Greater Los Angeles Healthcare System
Joe Thompson, MD, MPH Director, Arkansas Center for Health Improvement
Reginald L. Washington, MD, FAAP, FACC, FAHA, Professor of Pediatric Cardiology University of Colorado Medical Center

2016 Committee on Performance Measurement members:
Bruce Bagley, MD, American Medical Association
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, MPH, Medicalis
Helen Darling, National Quality Forum
Foster Gesten, MD, FACP, New York State Department of Health
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Group Health Physicians
Christine S. Hunter, MD (Co-chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services (DHHS)
Nancy Lane, PhD, Vanderbilt University Medical Center
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, PhD, RN, AARP Public Policy Institute
Eric C. Schneider, MD, MSc, FACP (Co-chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Blue Shield of California
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2008
Ad.3 **Month and Year of most recent revision:** 05, 2017
Ad.4 **What is your frequency for review/update of this measure?** Approximately every three years; sooner if the clinical guidelines change significantly
Ad.5 **When is the next scheduled review/update for this measure?** 2018

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. THE MEASURSE AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0034
Measure Title: Colorectal Cancer Screening (COL)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 50–75 years of age who had appropriate screening for colorectal cancer.
Developer Rationale: This measure encourages screening for colorectal cancer so that it can be prevented or detected early when it is most treatable, which reduces deaths associated with colorectal cancer.

Numerator Statement: Patients who received one or more screenings for colorectal cancer according to clinical guidelines.
Denominator Statement: Patients 51–75 years of age
Denominator Exclusions: This measure excludes patients with a history of colorectal cancer or total colectomy. The measure also excludes patients who use hospice services or are enrolled in an institutional special needs plan (SNP) or living long-term in an institution any time during the measurement year.

Measure Type: Process
Data Source: Claims, Electronic Health Data, Paper Medical Records
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** May 02, 2012

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary or **Summary of prior review in [year]**

- The 2017 United States Preventative Services Task Force (USPSTF) guidelines recommend screenings for colorectal cancer starting at age 50 and continuing until the age of 75. This guideline received an A rating, since the USPSTF concludes with high certainty that the benefits outweigh harms of performing colorectal cancer screening in patients age 50 to 75.
- The systematic review used to support this measure cites 47 articles (25 studies, fair or good quality) related to the effectiveness of screening programs based on the pre-specified screening tests (alone or in combination) in reducing incidence of and mortality from colorectal cancer; 44 articles (33 diagnostic accuracy studies, fair or good quality) related to the test performance characteristics of the pre-specified screening tests (alone or in combination) for detecting colorectal cancer, advanced adenomas, or adenomatous polyps based on size; and 113 articles (98 studies fair or good quality) related to the adverse effects of the different screening tests (either as single application or in a screening program) and variation in adverse effects by important subpopulations.
- The developer provides the following logic model for the measure: Adults at risk for colorectal cancer → Screening for colorectal cancer → Abnormal screening result → Evaluation and follow-up → Early detection and treatment of cancer → Improved length and/or quality of life

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer updated the Evidence form to provide the 2017 USPSTF guidelines (from the 2011 recommendation). Both the 2011 and 2017 guidelines recommend colorectal cancer screenings for individuals beginning at age 50 and continuing until age 75. Both guidelines received an A rating, meaning that the USPSTF recommends the service and there is high certainty that the net benefit is substantial.

Questions for the Committee:

- *The developer provided an updated 2017 USPSTF guideline to support the measure focus. Does the Committee agree that the measure reflects the current USPSTF recommendation?*

Guidance from the Evidence Algorithm

Measure a health outcome (Box 1) No → Assess performance of intermediate outcome, process, or structure(Box 3) Yes → Summary of QQC provided (Box 4) Yes → High certainty that the net benefit is substantial (Box 5) → High rating

Preliminary rating for evidence: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following performance rates from HEDIS, which reflects the most recent years of measurement:

Commercial Plans (HMO and PPO combined)			
Measurement Year	**2014**	**2015**	**2016**
Mean	61.2%	60.0%	60.1%
Std. dev.	8.9%	9.2%	9.6%

10th percentile	50.4%	49.2%	48.4%
25th percentile	54.9%	54.1%	53.9%
50th percentile	60.3%	59.5%	60.1%
75th percentile	67.6%	66.3%	66.4%
90th percentile	72.0%	71.6%	72.2%
Interquartile range	12.7	12.2	12.5

Medicare Rates (HMO and PPO combined)			
Measurement year	**2014**	**2015**	**2016**
Mean	65.5%	67.2%	67.7%
Std. dev.	11.6%	10.9%	12.4%
10th percentile	51.6%	52.6%	50.8%
25th percentile	59.9%	60.9%	60.9%
50th percentile	66.9%	68.1%	69.9%
75th percentile	73.1%	74.5%	76.4%
90th percentile	77.4%	79.6%	81.0%
Interquartile range	13.2	13.7	15.5

- From 2014 to 2016, performance rates for this measure have been generally stable or shown some improvement.
- The developer provided the following data for the denominator for the performance data for 2014, 2015, and 2016.

Commercial			
Measurement year	2014	2015	2016
Number of plans	401	415	412
Median denominator size by plan	411	411	411

Medicare			
Measurement year	2014	2015	2016
Number of plans	401	415	412
Median denominator size by plan	411	411	411

Disparities

- HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan.
- The developer provides disparities data from the literature, as follows:
    - Researchers have identified disparities in the rate of colorectal cancer screening based on race, ethnicity, income, education and English language proficiency. Racial/ethnic minorities, most notably Hispanic-Spanish, had lower colorectal cancer screening rates than Whites in 2010 (30.6% Hispanic-Spanish, 47.2% Asian, 49.5% American Indian/Alaska Native, 52.5% Hispanic-English, and 54.6% Native Hawaiian/Pacific Islander, compared to 62% White) (Liss and Baker, 2014).
    - Low-income populations have low colorectal cancer screening rates. The percentage of people who are up-to-date with screening has been consistently lower for people with a family income below 200 percent of the federal poverty level compared to people with a family income greater than or equal to

500 percent of the federal poverty level (In 2008, screening rate of 40.1% for people below 200 percent federal poverty level and 66.0% for people greater than or equal to 500 percent federal poverty level).

- The percentage of people who are up-to-date with screening has been consistently lower for people with less than a high school education compared to people with greater than a high school education (screening rate of 37.5% in less than high school and 62.0% in greater than high school). (Klabunde et al, 2011)
- Limited-English proficient populations exhibit lower colorectal cancer screening rates compared to English proficient populations. In 2006, 33% of Latinos responding in Spanish reported having been screened, compared to 51% of Latinos responding in English and 62% of English-speaking non-Latinos. (Diaz et al, 2008)

Citations

Diaz JA, Roberts MB, Goldman RE, Weitzen S, Eaton CB. Effect of language on colorectal cancer screening among latinos and non-latinos. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(8)2169-2173.

Klabunde CN, Cronin KA, Breen N, Waldron WR, Ambs AH, Nadel MR. Trends in colorectal cancer test use among vulnerable populations in the U.S. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2011;20(8):1611-1621.

Liss DT, Baker DW. Understanding current racial/ethnic disparities in colorectal cancer screening in the United States: the contribution of socioeconomic status and access to care. American Journal of Preventive Medicine. 2014;46(3):228-236.

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**
Evaluators: NQF Staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Staff Scientific Acceptability Preliminary Analysis

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Data are generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
- To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records, and registry). The developer anticipates that as electronic health records become more widespread the reliance on paper record review will decrease.

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- This measure is included in the composite Medicare Advantage Star Rating Program and is used in the Quality Payment Program (QPP).
- This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.
- This measure is used in the California P4P program, which is the largest non-governmental physician incentive program in the United States.
- This measure also is used in Quality Compass which is an tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance, as well as the NCQA Health Plan Ratings/Report Card. The measure is used in NCQA accreditation for both Health Plans and Accountable Care Organizations (ACO) as well as the Qualified Health Plan (QHP) Quality Rating System.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notations in medical record documentation are sufficient to meet measure criteria. Other questions have sought clarification about the screening methods that satisfy the measure numerator. During a recent public comment session, a majority of comments from measured entities supported updates to the measure to align with the latest clinical recommendations.
    - During the measure's last major update, feedback obtained through the NCQA feedback mechanisms resulted in specifications that include the new screening methods recommended by the USPSTF and other major clinical guideline organizations.

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- From 2014 to 2016, performance rates for this measure have been generally stable or shown improvement. In 2016, commercial plans on average performance rate of 60%, and Medicare plans had an average rate of 68%.
- Given the updated USPSTF guidelines for colorectal cancer screening and the recent changes to this measure, the developer believes performance may improve in the coming years. In 2016, two additional screening methods were added to the guideline and measure. The developer hypothesizes that addition of more screening options may help patients feel more comfortable with the screening process, and therefore increase the number of patients who choose to be screened for colorectal cancer.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer reported that no identified unintended consequences for this measure were identified during testing or since implementation.

Preliminary rating for Usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- 0658 : Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients (American Gastroenterological Association)
- Colorectal Cancer Screening – Minnesota Community Measurement (not NQF endorsed)

Harmonization

- The developer reports that the measure is harmonized to the extent possible.
- NQF #0658: Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients focuses on only one of the available screening methods: colonoscopy. The measure assesses whether patients who have had a colonoscopy also have a recommended follow-up interval of 10 years documented in their colonoscopy report, whereas NQF #0034 focuses on several available screening methods in addition to colonoscopy.
- The Minnesota Community Measurement quality measure is intended for use at the clinician or practice-level, whereas NQF#0034 is intended for use at the health plan level.

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Scientific Acceptability

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number: 0034
Measure Title: Colorectal Cancer Screening

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?
   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?
   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
   ☒Yes (go to Question #5)
   ☐No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☒Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
   Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☒High (go to Question #8)
   ☐Moderate (go to Question #8)
   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☐Yes (go to Question #9)
   ☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☐Yes (go to Question #10)

☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

    ☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?
   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.
   ☒Yes (go to Question #2)
   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?
   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?
   ☐Yes (please explain below then go to Question #3)
   ☐No (go to Question #3)
   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?
   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*
   ☐Yes (please explain below then go to Question #4)
   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
   ☐Yes (please explain below then go to Question #5)
   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
   ☐Yes (please explain below then go to Question #6)
   ☐No (go to Question #6)

☒Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☒Yes (go to Question #11)
    ☐No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☒Yes (go to Question #12)
    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #14)
☒Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

13. Was other validity testing reported?

☐Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

TIPS: Prior validity studies of the same data elements may be submitted

☐Yes (go to Question #15)
☒No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #16)
☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0034

Measure Title: Colorectal Cancer Screening

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 11/15/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in

such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

☐ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☒ Process: Colorectal cancer screening

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Adults at risk for colorectal cancer >>> Screening for colorectal cancer >>> Abnormal screening result >>> Evaluation and follow-up >>> Early detection and treatment of cancer >>> Improved length and/or quality of life

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☒ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2017 Submission** US Preventive Services Task Force (USPSTF). 2016. "Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement." *JAMA* 315(23):2564-2575. doi: 10.1001/jama.2016.5989 **2011 Submission** US Preventive Services Task Force (USPSTF). Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008 Nov 4;149(9):627-37. http://www.uspreventiveservicestaskforce.org/uspstf/uspscolo.htm
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**2017 Submission** "The USPSTF recommends screening for colorectal cancer starting at age 50 years and continuing until age 75 years (A recommendation)" **2011 Submission** The USPSTF recommends screening for colorectal cancer in adults, beginning at age 50 years and continuing until age 75 years. (A recommendation)

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2017 Submission** The USPSTF concludes with high certainty that the benefits outweigh harms of performing colorectal cancer screening in patients age 50 to 75.
Provide all other grades and definitions from the evidence grading system	**2017 Submission** N/A
Grade assigned to the **recommendation** with definition of the grade	**2017 Submission** Grade: A "The USPSTF recommends the service. There is high certainty that the net benefit is substantial." **2011 Submission** Grade: A The USPSTF recommends the service. There is high certainty that the net benefit is substantial.
Provide all other grades and definitions from the recommendation grading system	**2017 Submission** B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate, or there is moderate certainty that the net benefit is moderate to substantial. C: The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small. D: The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. I: The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined. **2011 Submission** B. The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. C. The USPSTF recommends against routinely providing the service. There may be considerations that support providing the service in an individual patient. There is at least moderate certainty that the net benefit is small.

	D. The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. I. The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2017 Submission** The evidence report supporting this guideline outlines the quantity and quality of evidence (Lin et al 2016). Key question 1: What is the effectiveness of screening programs based on the pre-specified screening tests (alone or in combination) in reducing incidence of and mortality from colorectal cancer? • Included 47 articles (25 studies, fair or good quality) Key question 2: What are the test performance characteristics of the prespecified screening tests (alone or in combination) for detecting colorectal cancer, advanced adenomas, or adenomatous polyps based on size? • Included 44 articles (33 diagnostic accuracy studies, fair or good quality) Key question 3a: What are the adverse effects of the different screening tests (either as single application or in a screening program)? Key Question 3b: Do adverse effects vary by important subpopulations (eg, age)? • Included 113 articles (98 studies, fair or good quality) Lin, J.S., M.A. Piper, L.A. Perdue, et al. 2016. "Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force." *JAMA* 315(23):2576-94. doi: 10.1001/jama.2016.3332. **2011 Submission**

	Quantity: Refer to USPSTF http://www.uspreventiveservicestaskforce.org/uspstf08/colocancer/coloartwhit.htm Quality: High
Estimates of benefit and consistency across studies	**2017 Submission** The USPSTF recommendation states: "The USPSTF concludes with high certainty that screening for colorectal cancer in average-risk, asymptomatic adults aged 50 to 75 years is of substantial net benefit. Multiple screening strategies are available to choose from, with different levels of evidence to support their effectiveness, as well as unique advantages and limitations, although there are no empirical data to demonstrate that any of the reviewed strategies provide a greater net benefit. Screening for colorectal cancer is a substantially underused preventive health strategy in the United States." **2011 Submission** The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.
What harms were identified?	**2017 Submission** The USPSTF guideline (2016) summarizes the harms of screening and early intervention: "The harms of screening for colorectal cancer in adults aged 50 to 75 years are small. The majority of harms result from the use of colonoscopy, either as the screening test or as follow-up for positive findings detected by other screening tests. The rate of serious adverse events from colorectal cancer screening increases with age."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2017 Submission** To our knowledge, there have been no published studies since the systematic review that would impact the recommendations.

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

0034_-_Colorectal_Cancer_Screening__-_Evidence_7.1.docx

1a.1 <u>For Maintenance of Endorsement:</u> Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

This measure encourages screening for colorectal cancer so that it can be prevented or detected early when it is most treatable, which reduces deaths associated with colorectal cancer.

1b.2. Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis. *(<u>This is required for maintenance of endorsement</u>. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data are stratified by year and product line (i.e. commercial, Medicare).

Colorectal Cancer Screening – commercial Rate (HMO and PPO Combined)

MEASUREMENT YEAR	MEAN	ST DEV	10TH	25TH	50TH	75TH	90TH	Interquartile Range
2014	61.2%	8.9%	50.4%	54.9%	60.3%	67.6%	72.0%	12.7
2015	60.0%	9.2%	49.2%	54.1%	59.5%	66.3%	71.6%	12.2
2016	60.1%	9.6%	48.4%	53.9%	60.1%	66.4%	72.2%	12.5

Colorectal Cancer Screening – Medicare Rate (HMO and PPO Combined)
MEASUREMENT YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2014 | 65.5% | 11.6% | 51.6% | 59.9% | 66.9% | 73.1% | 77.4% | 13.2
2015 | 67.2% | 10.9% | 52.6% | 60.9% | 68.1% | 74.5% | 79.6% | 13.7
2016 | 67.7% | 12.4% | 50.8% | 60.9% | 69.9% | 76.4% | 81.0% | 15.5

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 17.6 million Medicare beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the mean eligible population for the measure across health plans.

Colorectal Cancer Screening – commercial
YEAR | N Plans | Median Denominator Size per plan
2014 | 401 | 411
2015 | 415 | 411
2016 | 412 | 411

Colorectal Cancer Screening – Medicare
YEAR | N Plans | Median Denominator Size per plan
2014 | 449 | 396
2015 | 440 | 408
2016 | 459 | 411

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4
Researchers have identified disparities in the rate of colorectal cancer screening based on race, ethnicity, income, education and English language proficiency. Racial/ethnic minorities, most notably Hispanic-Spanish, had lower colorectal cancer screening rates than Whites in 2010 (30.6 percent Hispanic-Spanish, 47.2 percent Asian, 49.5 percent American Indian/Alaska Native, 52.5 percent Hispanic-English and 54.6 percent Native Hawaiian/Pacific Islander, compared to 62 percent White) (Liss and Baker, 2014). Low-income and low-literacy populations also have low colorectal cancer screening rates. The percentage of people who are up-to-date with screening has been consistently lower for people with a family income below 200 percent of the federal poverty level compared to people with a family income greater than or equal to 500 percent of the federal poverty level (In 2008, screening rate of 40.1 percent for people below 200 percent federal poverty level and 66.0 percent for people greater than or equal to 500 percent federal poverty level). Similarly, the percentage of people who are up-to-date with screening has been consistently lower

for people with less than a high school education compared to people with greater than a high school education (screening rate of 37.5 percent in less than high school and 62.0 percent in greater than high school). (Klabunde et al, 2011) Limited-English proficient populations exhibit lower colorectal cancer screening rates compared to English proficient populations. In 2006, 33 percent of Latinos responding in Spanish reported having a screen, compared to 51 percent of Latinos responding in English and 62 percent of English-speaking non-Latinos. (Diaz et al, 2008)

Brenner AT, Hoffman R, McWilliams A, Pignone MP, Rhyne RL, Tapp H, Weaver MA, Callan D, de Hernandez BU, Harbi K, Reuland DS. Colorectal cancer screening in vulnerable patients: promoting informed and shared decisions. American Journal of Preventive Medicine. 2016;51(4)454-462.

Diaz JA, Roberts MB, Goldman RE, Weitzen S, Eaton CB. Effect of language on colorectal cancer screening among latinos and non-latinos. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(8)2169-2173.

Klabunde CN, Cronin KA, Breen N, Waldron WR, Ambs AH, Nadel MR. Trends in colorectal cancer test use among vulnerable populations in the U.S. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2011;20(8):1611-1621.

Liss DT, Baker DW. Understanding current racial/ethnic disparities in colorectal cancer screening in the United States: the contribution of socioeconomic status and access to care. American Journal of Preventive Medicine. 2014;46(3):228-236.

Rice K, Gressard L, DeGroff A, Gersten J, Robie, J, Leadbetter S, Glover-Kudon R, Butterly L. Increasing colonoscopy screening in disparate populations: results from an evaluation of patient navigation in the New Hampshire Colorectal Cancer Screening Program. Cancer. 2017;123(17)3356-3366.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cancer : Colorectal

De.6. **Non-Condition Specific***(check all the areas that apply):*
Primary Prevention

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Elderly, Populations at Risk : Dual eligible beneficiaries

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
N/A

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment **Attachment:** 0034_COL_Value_Sets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
Yes

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Since the last NQF review, two additional screening methods have been added to the measure, in alignment with updates to clinical guidelines. These changes were reviewed by stakeholder groups, vetted through a public comment period, and approved by our committees.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients who received one or more screenings for colorectal cancer according to clinical guidelines.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
ADMINISTRATIVE:
Patients who received one or more screenings for colorectal cancer. Any of the following meet criteria:
-Fecal occult blood test (FOBT Value Set) during the measurement year.
-Flexible sigmoidoscopy (Flexible Sigmoidoscopy Value Set) during the measurement year or the four years prior to the measurement year.
-Colonoscopy (Colonoscopy Value Set) during the measurement year or the nine years prior to the measurement year.
-CT colonography (CT Colonography Value Set) during the measurement year or the four years prior to the measurement year.
-FIT-DNA test (FIT-DNA Value Set) during the measurement year or the two years prior to the measurement year.

MEDICAL RECORD:
Patients who received one or more screenings for colorectal cancer. Any of the following meet criteria:
-Fecal occult blood test during the measurement year.
-Flexible sigmoidoscopy during the measurement year or the four years prior to the measurement year.
-Colonoscopy during the measurement year or the nine years prior to the measurement year.
-CT colonography during the measurement year or the four years prior to the measurement year.
-FIT-DNA test during the measurement year or the two years prior to the measurement year.

Documentation in the medical record must include a note indicating the date when the colorectal cancer screening was performed. A result is not required if the documentation is clearly part of the "medical history" section of the record; if this is not clear, the result or finding must also be present (this ensures that the screening was performed and not merely ordered).

A pathology report that indicates the type of screening (e.g., colonoscopy, flexible sigmoidoscopy) and the date when the screening was performed meets criteria.
For pathology reports that do not indicate the type of screening and for incomplete procedures:

--Evidence that the scope advanced beyond the splenic flexure meets criteria for a completed colonoscopy.
--Evidence that the scope advanced into the sigmoid colon meets criteria for a completed flexible sigmoidoscopy.

There are two types of FOBT tests: guaiac (gFOBT) and immunochemical (FIT). Depending on the type of FOBT test, a certain number of samples are required for numerator compliance. Follow the instructions below to determine member compliance.
--If the medical record does not indicate the type of test and there is no indication of how many samples were returned, assume the required number was returned. The member meets the screening criteria for inclusion in the numerator.
--If the medical record does not indicate the type of test and the number of returned samples is specified, the member meets the screening criteria only if the number of samples specified is greater than or equal to three samples. If there are fewer than three samples, the member does not meet the screening criteria for inclusion.
--FIT tests may require fewer than three samples. If the medical record indicates that an FIT was done, the member meets the screening criteria, regardless of how many samples were returned.
--If the medical record indicates that a gFOBT was done, follow the scenarios below.
–If the medical record does not indicate the number of returned samples, assume the required number was returned. The member meets the screening criteria for inclusion in the numerator.
–If the medical record indicates that three or more samples were returned, the member meets the screening criteria for inclusion in the numerator.
–If the medical record indicates that fewer than three samples were returned, the member does not meet the screening criteria.

Do not count digital rectal exams (DRE), FOBT tests performed in an office setting or performed on a sample collected via DRE.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 51–75 years of age

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients 51–75 years of age as of the end of the measurement year (e.g. December 31).

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
This measure excludes patients with a history of colorectal cancer or total colectomy. The measure also excludes patients who use hospice services or are enrolled in an institutional special needs plan (SNP) or living long-term in an institution any time during the measurement year.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
Exclude patients with either of the following any time during the patient's history through December 31 of the measurement year:
- Colorectal cancer (Colorectal Cancer Value Set)
- Total colectomy (Total Colectomy Value Set)

Exclude patients who use hospice services any time during the measurement year (Hospice Value Set).

Exclude patients 65 and older who are enrolled in an institutional SNP or living long-term in an institution at any time during the measurement year.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
None

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification

If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Step 1. Determine the eligible population: identify patients 51-75 years of age by the end of the measurement year.
Step 2. Search for an exclusion in the patient's history: history of total colectomy or colorectal cancer. Exclude these patients from the eligible population.
Step 3. Determine numerator: the number of patients who have been screened for colorectal cancer by any of the included screening methods, within the associated time interval.
Step 4. Calculate the rate.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Data, Paper Medical Records

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2. Validity – See attached Measure Testing Submission Form
0034_-_Colorectal_Cancer_Screening__-_Testing_7.1-636463498807302646.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 0034
Measure Title: Colorectal Cancer Screening
Date of Submission: 11/15/2017
Type of Measure:

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

Instructions

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

- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed 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]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

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

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? 2017 Submission: 2016 2011 Submission: 2009

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

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

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*)

2017 Submission

Sample for measure score reliability testing: The measure score reliability was calculated from HEDIS data that included 459 Medicare health plans and 412 commercial health plans. The sample included all Medicare and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Sample for Construct Validity Testing: Construct validity was calculated from HEDIS data that included 430 Medicare health plans and 412 commercial health plans. The sample included all Medicare and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

2011 Submission

HEDIS Health Plan performance data 2010

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

2017 Submission

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

Product Type	Number of Plans	Median number of eligible patients per plan
Commercial	412	411
Medicare	459	411

Beneficiary Sample for Construct Validity Testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 17.6 million Medicare beneficiaries. Data is summarized at the health plan level. Data are stratified by product line (i.e. commercial, Medicare). Below is a description of the sample.

It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of plans	Median number of eligible patients per plan
Commercial	412	411
Medicare	430	411

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

Validity was demonstrated through construct validity.

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.
For Medicare health plans, this measure was analyzed by low-income status, dual eligibility and disability, which served as proxies for lower socioeconomic status. These are available data elements for Medicare 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*)
2017 Submission
Reliability Testing of Performance Measure Score: same as below

2011 Submission
Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the

confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

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

2017 Submission

Beta-Binomial Statistic:

Commercial	Medicare
0.997	0.988

2011 Submission

Commercial Plans 2010: reliability 0.994468
Medicaid 2010: Not available
Medicare 2010: reliability 0.993543

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?*)

2017 Submission

Interpretation of measure score reliability testing: The testing suggests the measure has high reliability.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

2017 Submission

We assessed face validity and construct validity for this measure.

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

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs – whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness, and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure “Desirable”? The work-up is vetted by NCQA’s Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in

clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

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

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

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

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review, and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation, information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.
Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the new year's HEDIS Volume 2.

Method of testing construct validity: We tested for construct validity by exploring whether Colorectal Cancer Screening was correlated with Breast Cancer Screening. We hypothesized that organizations that perform well on Colorectal Cancer Screening should perform well on Breast Cancer Screening. To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

2011 Submission

NCQA tested the measure for face validity using a panel of stakeholders with specific expertise in measurement. This panel included representatives from key stake holder groups, including oncologists, family practitioners, health plans, state Medicaid agencies and researchers. Experts reviewed the results of the field test and assessed whether the results were consistent with expectation, whether the measure represented quality care, and whether we were measuring the most important aspects of care in this area.

In the pilot test, we explored periodicities associated with colorectal cancer screening, as long periodicities in light of average lengths of enrollment in MCOs can be a threat to validity. We examined whether the rates of screening would differ depending on the length of time an individual had been enrolled in the plan and found little effect as shown in Table 2. Although the rates increase a small amount each year in each plan, the relative rates of screening remain about the same. The sample sizes decline significantly with increased lengths of continuous enrollment; at 10 years, only two MCOs had enough data to estimate the rate.

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

2017 Submission

Results of face validity assessment:
Input from our multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity.

Statistical results of construct validity testing: The results in Table 1a and Table 1b indicate that there is a strong, positive relationship between the Colorectal Cancer Screening measure and the Breast Cancer Screening measure. This relationship is statistically significant ($p<0.0001$).

Table 1a. Correlations in Commercial Measures – 2016

	Pearson Correlation Coefficient
	Breast Cancer Screening
Colorectal Cancer Screening	0.711

Note: $p<0.0001$

Table 1b. Correlations in Medicare Measures – 2016

	Pearson Correlation Coefficient
	Breast Cancer Screening
Colorectal Cancer Screening	0.716

Note: $p<0.0001$

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

2017 Submission

Interpretation of systematic assessment of face validity: These results indicate the technical expert panel showed good agreement that the measures as specified will accurately differentiate quality across providers. Our interpretation of these results is that this measure has sufficient face validity.

Interpretation of construct validity testing: The two measures had high correlation, which indicates the measure has good construct validity.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b3*

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

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

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

2b3.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with risk factors**
☐ **Stratification by risk categories**
☐ **Other,**

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

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

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:
☐ **Published literature**
☐ **Internal data analysis**
☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of*

unique variation in the outcome, assessment of between-unit effects and within-unit effects.) **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

2017 Submission

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans' performance is significantly different from each other.

2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2017 Submission

HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Com.	8582	60.1	9.6	48.4	53.9	60.1	66.4	72.2	12.5	<0.001
Medicare	1330	67.7	12.4	50.8	60.9	69.9	76.4	81.0	15.5	<0.001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2017 Submission

The difference between the 25th and 75th percentile is statistically significant for both product lines. For commercial plans, there is a 12.5 percentage point gap between 25th and 75th percentile plans. This gap represents an average 1,073 more patients that have been screened for colorectal cancer compared to low performing plans (estimated from average health plan eligible population).

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

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

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

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

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2017 Submission

This measure is collected with a complete sample.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

2017 Submission

This measure is collected with a complete sample.

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

2017 Submission

This measure is collected with a complete sample.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records, and registry). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO's ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) Information practices and control procedures
2) Sampling methods and procedures
3) Data integrity
4) Compliance with HEDIS specifications
5) Analytic file production
6) Reporting and documentation

In addition to the HEDIS audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system, NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system informs both annual updates to the measures as well as routine re-evaluation of measures. These processes include updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Broad public use and dissemination of these measures are encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Quality Improvement (Internal to the specific organization)	Payment Program Medicare STARS https://www.medicare.gov/find-a-plan/questions/home.aspx California´s Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p Quality Payment Program https://qpp.cms.gov Regulatory and Accreditation Programs Accreditation http://www.ncqa.org/Programs/Accreditation/Health-Plan-HP.aspx HEDIS ACO http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx Quality Improvement (external benchmarking to organizations) Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

CMS MEDICARE ADVANTAGE STAR RATING PROGRAM: This measure is included in the composite Medicare Advantage Star Rating. CMS calculates a Star Rating (1-5) for all Medicare Advantage health plans based on 53 performance measures. Medicare beneficiaries can view the star rating and individual measure scores on the CMS Plan Compare website. The Star Rating is also used to calculate bonus payments to health plans with excellent performance. The Medicare Advantage Plan Rating program covers 11.5 million Medicare beneficiaries in 455 health plans across all 50 states.

NCQA STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012, the report included measures on 11.5 Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2012, a total of 455 Medicare Advantage health plans, 404 commercial health plans, and 136 Medicaid health plans across 50 states were included in the rankings.

NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans. In 2012, a total of 170 Medicare Advantage health plans were accredited using this measure among others covering 7.1 million Medicare beneficiaries and 336 commercial health plans covering 87 million lives. Health plans are scored based on performance compared to benchmarks.

QUALIFIED HEALTH PLAN (QHP) QUALITY RATING SYSTEM (QRS): This measure is used in the Qualified Health Plan (QHP) Quality Rating System, which provides comparable information to consumers about the quality of health care services and QHP enrollee experience offered in the Marketplaces.

NCQA ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

CALIFORNIA VALUE BASED PAY FOR PERFORMANCE PROGRAM: This measure is used in the California P4P program, which is the largest non-governmental physician incentive program in the United States. Founded in 2001, it is managed by the Integrated Healthcare Association (IHA) on behalf of ten health plans representing 9 million insured persons. IHA reports results on approximately 35,000 physicians in 200 physician organizations.

QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.

NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notation in medical record documentation is sufficient to meet measure criteria. Other questions have sought clarification about the screening methods that satisfy the measure numerator. During a recent public comment session, a majority of comments from measured entities supported updates to the measure to align with the latest clinical recommendations.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the Medicare Advantage Star Rating program.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
During the measure's last major update, feedback obtained through the mechanisms described in 4a2.2.1 informed how we revised the measure to include new screening methods recommended by the U.S. Preventive Services Task Force and other major clinical guideline organizations.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
From 2014 to 2016, performance rates for this measure have been generally stable or shown slight improvement. In 2016, commercial plans on average performance rate of 60 percent, and Medicare plans had an average rate of 68 percent. There continues to be significant variation between the 10th and 90th percentiles, suggesting room for improvement. In 2016, commercial plans in the 10th percentile had a rate of 48 percent, compared to 72 percent among plans in the 90th percentile. For Medicare, plans in the 10th percentile had a rate of 51 percent compared to 81 percent among plans in the 90th percentile.

Given the new US Preventive Services Task Force guidelines for colorectal cancer screening and our recent changes to this measure, we may see performance improvement in the coming years. In 2016, two additional screening methods were added to the guideline and measure. The addition of more screening options may help patients feel more comfortable with the screening process, and therefore increase the number of patients who choose to be screened for colorectal cancer.

4b2. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
There were no identified unintended consequences for this measure during testing or since implementation.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits for this measure during testing or since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0658 : Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
Colorectal Cancer Screening – Minnesota Community Measurement

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Minnesota Community Measurement: These measures are harmonized but intended for different levels of accountability. --Both measures exclude patients who have had a total colectomy, a history of colorectal cancer, or who have been in hospice care. --Both measures include the same screening methods and intervals. --The Minnesota Community Measurement quality measure is intended for use at the clinician or practice-level, whereas NQF#0034 is intended for use at the health plan level. American Gastroenterological Association: These measures have different areas of focus and are harmonized where appropriate. --The American Gastroenterological Association measure focuses on only one of the available screening methods: colonoscopy. The measure assesses whether patients who have had a colonoscopy also have a recommended follow-up interval of 10 years documented in their colonoscopy report.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The NCQA Colorectal Cancer Screening Measurement Advisory Panels advised NCQA during measure development. They evaluated the way staff specified the measure, reviewed field test results, and assessed NCQA's overall desirable attributes of Relevance, Scientific Soundness, and Feasibility. The advisory panel consisted of a balanced group of experts. In addition to this advisory panel, we vetted the measure with a host of other stakeholders, as is our process. Thus, our measures are the result of consensus from a broad and diverse group of stakeholders.

2008 Colorectal Cancer Measurement Advisory Panel members:
Joel V. Brill, Predictive Health, LLC
Durado Brooks, American Cancer Society
Robert Fletcher, Harvard Medical School
William Lawrence, AHRQ Center for Outcomes and Effectiveness
T.R. Levin, Kaiser Permanente
Michael Pignone, UNC Hospital
Evelyn Whitlock

2016 Colorectal Cancer Screening Measurement Advisory Panel members:
Matthew Barish, MD FACR, Stony Brook University Hospital
Linda Berthold, PhD, Central California Alliance for Health
Durado Brooks, MD MPH, American Cancer Society
Joseph Chin, MD MS, Centers for Medicare and Medicaid Services
T.R. Levin, MD, Kaiser Permanente Northern California
Steven Phillips, MD CMD, Sierra Health Services Inc
Tim Wilt, MD MPH, VA Medical Center Minneapolis
Ann Zauber, PhD, Memorial Sloan Kettering Cancer Center

2016 Geriatric Measurement Advisory Panel members:
Wade Aubry, UCSF Institute for Health Policy Studies
Arlene Bierman, AHRQ
Patricia A. Bomba, MD FACP, Excellus BlueCross BlueShield
Jennie Chin Hansen, RN, American Geriatrics Society
Joyce Dubow, Public Member/Consumer Advocate
Peter Hollman, Brown University
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Steven Phillips, MD, CMD, Geriatric Specialty Care
Eric G. Tangalos, MD, FACP, AGSF, CMD, Mayo Clinic
Dirk Wales, MD, PsyD, Cigna HealthSpring
Joan Weiss, PhD, RN, CRNP, U.S. Department of Health and Human Services
Neil Wenger, MD, UCLA Division of Medicine

2016 Committee on Performance Measurement members:
Bruce Bagley, MD, American Medical Association
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, MPH, Medicalis
Helen Darling, National Quality Forum
Foster Gesten, MD, FACP, New York State Department of Health
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services

David Grossman, MD, MPH, Group Health Physicians
Christine S. Hunter, MD (Co-chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services (DHHS)
Nancy Lane, PhD, Vanderbilt University Medical Center
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, PhD, RN, AARP Public Policy Institute
Eric C. Schneider, MD, MSc, FACP (Co-chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Blue Shield of California
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms

2016 Technical Measurement Advisory Panel members:
Andy Amster, MSPH, Kaiser Permanente
Jennifer Brudnicki, MBA, Geisinger Health Plan
Lindsay Cogan, PhD, MS, New York State Department of Health
Kathy Coltin, MPH, Independent Consultant
Mike Farina, MVP Healthcare
Marissa Finn, MBA, CIGNA HealthCare
Scott Fox, MS, Med,Independence Blue Cross
Carlos Hernandez, CenCal Health
Harmon Jordan, ScD, RTI International
Virginia Raney
Lynne Rothney-Kozlak, MPH, Rothney-Kozlak Consulting, LLC
Laurie Spoll, Aetna
Natan Szapiro, Independent Consultant

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2003

Ad.3 **Month and Year of most recent revision:** 10, 2016

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.

Ad.5 **When is the next scheduled review/update for this measure?** 2018

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. THE MEASURES AND SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, “commercial use” refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2508
Measure Title: Prevention: Dental Sealants for 6-9 Year-Old Children at Elevated Caries Risk, Dental Services
Measure Steward: American Dental Association on behalf of the Dental Quality Alliance
Brief Description of Measure: Percentage of enrolled children in the age category of 6-9 years at "elevated" risk (i.e., "moderate" or "high") who received a sealant on a permanent first molar tooth within the reporting year.
Developer Rationale: Inequalities in oral health status and inadequate use of oral health care services are well documented. Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, L i, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009). Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions.

Evidence-based clinical recommendations recommend that sealants be placed on pits and fissures of children's primary and permanent teeth when it is determined that the tooth, or the patient, is at risk of experiencing caries (Beauchamp et al. 2008). The evidence for sealant effectiveness in permanent molars is stronger than evidence for primary molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013).

The proposed measure, Prevention: Sealants for 6-9 Year-Old Children at Elevated Caries Risk, captures whether children at moderate or high caries risk received a sealant on a permanent first molar tooth. Permanent first molars usually erupt between ages 6 and 7 years. Thus, this measure addresses both the tooth type on which sealants are placed and the timeliness of care provision. The measure Sealants for 6-9 Year-Old Children allows plans and programs to assess whether children at risk for caries are receiving evidence-based prevention and target performance improvement initiatives accordingly.

This measure is a program/plan specific measure that contributes to the Healthy People 2020 Objective OH 12.2 that calls for increasing the percent children aged 6 to 9 years who received dental sealants on one or more of their first permanent molars.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

Numerator Statement: Unduplicated number of enrolled children age 6-9 years at "elevated" risk (i.e., "moderate" or "high") who received a sealant on a permanent first molar tooth as a dental service.
Denominator Statement: Unduplicated number of enrolled children age 6-9 years who are at "elevated" risk (i.e., "moderate" or "high")
Denominator Exclusions: Medicaid/ CHIP programs should exclude those individuals who do not qualify for dental benefits. The

exclusion criteria should be reported along with the number and percentage of members excluded.

There are no other exclusions.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Sep 18, 2014 **Most Recent Endorsement Date:** Sep 18, 2014

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- Sealants for 6-9 Year-Old Children at Elevated Caries Risk indicates the percentage of children at moderate to high risk for caries who received a sealant on a first permanent molar. Evidence-based clinical recommendations recommend that sealants be placed on pits and fissures of children's primary and permanent teeth when it is determined that the tooth, or the patient, is at risk of experiencing caries, with greater evidence of effectiveness in permanent molars compared to primary molars (Beauchamp et al. 2008).
- Grade/Strength of Recommendation: B which is defined as: "Directly based on category II evidence or extrapolated recommendation for category I evidence."

Citation:
Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268. Available at: http://jada.ada.org/content/139/3/257.full.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☒ **The developer provided updated evidence for this measure:**

Updates:

- A recent Cochrane Review on the effectiveness of sealants brings together all the evidence on this topic. The conclusions of this new review continue to support the recommendations of the ADA Sealant Guideline (Note: the ADA is currently updating this guideline).

- Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review, but does note a recent Cochrane review collated all evidence and reached the same conclusions that supported the original guideline. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3)→ QQC presented (Box 4)→ Contains Quantity: High (7 systematic reviews and 14 individual clinical studies) Quality: High, Consistency: Moderate → Rate as High

Preliminary rating for evidence: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The Developer used data from five sources and refer to "program" level information and "plan" level information (Texas Medicaid, Texas CHIP, Florida CHIP, and Florida Medicaid programs as well as national commercial data from Dental Service of Massachusetts, Inc.). The developer presented the total number of children enrolled in each program/plan. In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011.
- The data source and sample size are sufficient to assess gaps in performance. The performance range of **20% to 30% in CY 2010** (year in which data were available for all five programs) indicate variation in sealant replacement across programs. Data from the Centers for Medicare and Medicaid Services (CMS) indicate significant variation among state Medicaid programs, ranging from 6% to 31% of children 6-9 years old, who received a sealant on a permanent molar tooth (CMS-416 data, FY 2011).
- The developer did not provide more recent performance data, stating that due to the start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting.

Disparities

- Disparities by geographic location were detected for two programs. Statistically significant difference in performance by race and ethnicity also were detected in the two programs for which there were race/ethnicity data. In addition, the developers also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), Medicaid program type, CHIP dental plan, commercial product line, and preferred language for program communications. The developers detected disparities based on each of these various factors, but data on all of these characteristics were not consistently available for all programs so they presented disparities data on those characteristics that were most consistently available and had the greatest standardization (i.e. race/ethnicity and geographic location).

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.
2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.
Validity
2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.
2b2-2b6. Potential threats to validity should be assessed/addressed.

Staff Scientific Acceptability Rating Logic

*The original testing was submit as permitted by NQF.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure relies on **standard data elements in administrative claims data** (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes.

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details
- Texas Health and Human Services Commission: Texas Medicaid/CHIP https://hhs.texas.gov/sites/default/files//documents/lawsregulations/ handbooks/umcm/6-2-15.pdf

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- This measure is included in the CHIPRA Core Measures Program. Some Medicaid programs noted that they do not receive complete data on tooth number from their contracted plans, which is a required data element for this measure. As a result, the affected programs must manually get these data from their contracted plans. Because tooth number is required for reimbursement, these data are readily accessible for plan level reporting. Despite initial concerns about this data element, 25 states reported this measure in FY 2015, and 34 reported in FY 2016.
- A dental benefits administrator (DBA) suggested that the DQA consider adding patient exclusions to the measure. The DQA considered exclusions previously during initial measure development and during annual reviews. Exclusions were not incorporated due to concerns about the introduction of biased measurement, increasing measurement complexity, and adversely affecting implementation feasibility. However, the DQA continues to monitor this issue and will revisit it during the 2018 annual review. The DQA has invited the DBA to present its suggestion with supporting data to the DQA. The DQA also has invited other DBAs and Medicaid program administrators to provide input. All of this stakeholder feedback will be incorporated into the next annual review.

Additional Feedback:

- This measure was one of 10 performance measures approved by the Dental Quality Alliance (DQA) that focused on Dental Caries Prevention and Disease Management among children. The Dental Quality Alliance (DQA) was formed at the request of 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 the 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.

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

The developer provides initial reporting data available from the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after approval by the DQA and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 461207, 25.41, 25.59, 25.53
2015, 503515, 24.99, 25.18, 24.91

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 76415, 20.17, 22.30, 21.69
2015, 58833, 20.20, 23.14, 22.43

The developer notes that these data also suggest fairly stable rates over the two-year period (i.e. improvement is not noted). However, as noted above, these are initial performance data; additional time may be needed to see improvement within this program because most measure users are just now getting their quality measurement programs underway.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation [unexpected findings]

No unintended or negative consequences were identified by the developer.

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- N/A

Harmonization

- N/A

Committee pre-evaluation comments
Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Staff Scientific Acceptability Rating Logic

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?
   ☒Yes (use your rating from data element validity testing – Question #16- under Validity Section)
   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?
   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
   ☐Yes (go to Question #5)
   ☒No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☐Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

6. **RATING (score level) -** What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
   Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☐High (go to Question #8)
   ☐Moderate (go to Question #8)
   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☒Yes (go to Question #9)
   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #10)
☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?
   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.
   ☒Yes (go to Question #2)
   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?
   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?
   ☐Yes (please explain below then go to Question #3)
   ☐No (go to Question #3)
   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?
   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*
   ☐Yes (please explain below then go to Question #4)
   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
   ☐Yes (please explain below then go to Question #5)
   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
   ☐Yes (please explain below then go to Question #6)
   ☒No (go to Question #6)

☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☐Yes (go to Question #11)
    ☒No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☐Yes (go to Question #12)
    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #14)
☐Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

13. Was other validity testing reported?

☒Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

TIPS: Prior validity studies of the same data elements may be submitted

☒Yes (go to Question #15)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #16)
☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Prevention: Dental Sealants for 6-9 Year-Old Children at Elevated Caries Risk

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 2/10/2014

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☐ Health outcome:

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome:

X☐ Process: Receipt of evidence-based preventive dental service – sealants on permanent molars - during the reporting period

☐ Structure:

☐ Other:

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not applicable.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

Not applicable.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Sealants for 6-9 Year-Old Children at Elevated Caries Risk indicates the percentage of children at moderate to high risk for caries who received a sealant on a first permanent molar. Evidence-based clinical recommendations recommend that sealants be placed on pits and fissures of children's primary and permanent teeth when it is determined that the tooth, or the patient, is at risk of experiencing caries, with greater evidence of effectiveness in permanent molars compared to primary molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013). This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (sealants) as well as the specific tooth type for which the evidence is the strongest (permanent molar) and the timing of sealant placement to maximize effectiveness (shortly after eruption – 6-9 years of age for permanent first molars). As described in 1b1 (Importance), dental caries is the most common chronic disease in children in the U.S. and a significant percentage of children have untreated dental caries. Dental decay causes significant short- and long-term adverse consequences for children's health and functioning. As detailed below, timely placement of sealants on permanent first molars have demonstrated effectiveness in reducing caries among children, thereby improving oral health, overall health, and overall well-being.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268. Available at: http://jada.ada.org/content/139/3/257.full.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

"Caries Prevention: Sealants should be placed on pits and fissures of **children's** and **adolescents'** permanent teeth when it is determined that the tooth, or the patient, is at risk of developing caries." (Beauchamp et al. 2008, p. 263, Table 3)

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade/Strength of Recommendation: B which is defined as: "Directly based on category II evidence or extrapolated recommendation for category I evidence." (Beauchamp 2008, pp. 261, 263, Tables 1, 2, 3)
[See grades for strength of evidence in section 1a7.]

Grading system adapted from: Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. BMJ 1999;318(7183):593-596.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

A: Directly based on category I evidence

B: Directly based on category II evidence or extrapolated recommendation from category I evidence

C: Directly based on category III evidence or extrapolated recommendation from category I or II evidence

D: Directly based on category IV evidence or extrapolated recommendation from category I, II or III evidence

Grading system adapted from: Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. BMJ 1999;318(7183):593-596.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same as that provided for the guidelines provided in 1a.4.1.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ **X**Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable.

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

Not applicable.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable.

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Not applicable.

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable.

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268. Available at: http://jada.ada.org/content/139/3/257.full.

Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Not applicable.

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The following four clinical questions were addressed:

- "Under what circumstances should sealants be placed to prevent caries?"

- "Does placing sealants over early (noncavitated) lesions prevent progression of the lesions?"
- "Are there conditions that favor the placement of resin-based versus glass ionomer cement sealants in terms of retention or caries prevention?"
- "Are there any techniques that could improve sealants' retention and effectiveness in caries prevention?" (Beauchamp et al. 2008, pp. 259-260)

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

"Caries Prevention: Sealants should be placed on pits and fissures of **children's** and **adolescents'** permanent teeth when it is determined that the tooth, or the patient, is at risk of developing caries." (Beauchamp et al. 2008, p. 263, Table 3)

Grade: The evidence grade is **IA** which is defined as: "Evidence from systematic reviews of randomized controlled trials" (Beauchamp 2008, pp. 261, 263, Tables 1, 3). Grading system adapted from: Shekelle et al. (1999) cited in 1a.4.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Ia: Evidence from systematic reviews of randomized controlled trials

Ib: Evidence from at least one randomized controlled trial

IIa: Evidence from at least one controlled study without randomization

IIb: Evidence from at least one other type of quasiexperimental study, such as time series analysis or studies in which the unit of analysis is not the individual

III: Evidence from nonexperimental descriptive studies, such as comparative studies, correlation studies, cohort studies and case-control studies

IV: Evidence from expert committee reports or opinions or clinical experience of respected authorities

(Beauchamp et al. 2008, p. 261) Grading system adapted from: Shekelle et al. (1999).

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).

Date range: Literature studies for sealants were conducted to identify all systematic reviews through Oct. 4, 2006. To ensure new clinical studies published since the search within each review were included within the guideline development effort, additional searches were conducted for clinical trials until September 2006.

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

7 systematic reviews and 14 individual clinical studies were reviewed with respect to the clinical questions identified. The evidence guidelines do not provide summary data regarding the number of studies by type of study. (Beauchamp 2008, p. 260)

However, the guidelines provide the following details regarding the strength and quality of the evidence related to sealants for caries prevention:

Evidence Grade Ia (systematic reviews of randomized controlled trials)

Supports the following evidence statements based on the evidence review by the expert panel:

- "Placement of resin-based sealants on the permanent molars of children and adolescents is effective for caries reduction." (Beauchamp 2008, p. 260)
- "Reduction of caries incidence in children and adolescents after placement of resin-based sealants ranges from 86 percent at one year to 78.6 percent at two years and 58.6 percent at four years." (Beauchamp 2008, p. 260)

Studies with evidence grade of Ia cited:

Ahovuo-Saloranta A, Hiiri A, Nordblad A, Worthington H, Mäkelä M. Pit and fissure sealants for preventing dental decay in the permanent teeth of children and adolescents. Cochrane Database Syst Rev 2004(3):CD001830.

Llodra JC, Bravo M, Delgado-Rodriguez M, Baca P, Galvez R. Factors influencing the effectiveness of sealants: a meta-analysis. Community Dent Oral Epidemiol 1993;21(5):261-268.

Evidence Grade Ib (evidence from at least one randomized controlled trial)

Supports the following evidence statements based on the evidence review by the expert panel:

- "Sealants are effective in reducing occlusal caries incidence in permanent first molars of children, with caries reductions of 76.3 percent at four years, when sealants were reapplied as needed. Caries reduction was 65 percent at nine years from initial treatment, with no reapplication during the last five years." (Beauchamp 2008, p. 261)

Studies with evidence grade of Ib cited:

Bravo M, Montero J, Bravo JJ, Baca P, Llodra JC. Sealant and fluoride varnish in caries: a randomized trial. J Dent Res 2005;84(12):1138-1143.

Evidence Grade III (evidence from nonexperimental descriptive studies, such as comparative studies, correlation studies, cohort studies and case control studies)

Supports the following evidence statements based on the evidence review by the expert panel:

- "There is consistent evidence from private dental insurance and Medicaid databases that placement of sealants on first and second permanent molars in children and adolescents is associated with reductions in the subsequent provision of restorative service." (Beauchamp 2008, p. 261)
- "Evidence from Medicaid claims data for children who were continuously enrolled for four years indicates that sealed permanent molars are less likely to receive restorative treatment, that the time between receiving sealants and receiving restorative treatment is greater, and that the restorations were less extensive than those in permanent molars that were unsealed." (Beauchamp 2008, p. 261)

Studies with evidence grade of III cited:

Bhuridej P, Damiano PC, Kuthy RA, et al. Natural history of treatment outcomes of permanent first molars: a study of sealant effectiveness. JADA 2005;136(9):1265-1272.

Dennison JB, Straffon LH, Smith RC. Effectiveness of sealant treatment over five years in an insured population. JADA 2000;131(5):597-605.

Hotuman E, Rølling I, Poulsen S. Fissure sealants in a group of 3-4-year-old children. Int J Paediatr Dent 1998;8(2):159-160.

Weintraub JA, Stearns SC, Rozier RG, Huang CC. Treatment outcomes and costs of dental sealants among children enrolled in Medicaid. Am J Public Health 2001;91(11):1877-1881.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The quality of the evidence is high, grades of Ia (systematic reviews of randomized controlled trials), for sealants placed on permanent molars of children and adolescents.

The evidence directly pertains to both the measure focus and the measure target population.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Meta-analyses were not conducted as part of the evidence review. Please see the response in 1a.7.5. regarding the identified benefits and associated strength of evidence. However, a more recent Cochrane Review published in 2013 by Ahovuo-Saloranta et al. brings together all the evidence in a quantitative manner. More information from this review is provided below in Section 1.a.7.9

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Harms were not evaluated as part of this systematic review. However this question was addressed in a recent Cochrane Review on the effectiveness of sealants (Ahovuo-Saloranta et al. 2013), and it was noted: "Only two studies (Bravo 2005; Liu 2012) assessed side effects of the sealants. No adverse effects were detected or reported by patients included in the studies."

Citations:

Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

Bravo M, Montero J, Bravo JJ, Baca P, Llodra JC. Sealant and fluoride varnish in caries: a randomized trial. Journal of Dental Research 2005;84(12):1138-43.

Liu BY, Lo ECM, Chu CH, Lin HC. Randomized trial on fluorides and sealants for fissure caries prevention. Journal of Dental Research 2012;91(8):753-8.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

A recent Cochrane Review on the effectiveness of sealants brings together all the evidence on this topic. The conclusions of this new review continue to support the recommendations of the ADA Sealant Guideline (Note: the ADA is currently updating this guideline). The summary of findings from the Cochrane review appears below:

SUMMARY OF FINDINGS FOR THE MAIN COMPARISON *[Explanation]*

Resin-based sealant compared to control without sealant for preventing dental caries

Patient or population: Children and adolescents
Settings: Sealant applications for school children in USA, Canada, China & Colombia
Intervention: Resin-based sealant applications on occlusal tooth surfaces of permanent molars
Comparison: No sealant application

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	Number of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk	Corresponding risk				
	Control teeth	Sealed teeth				
Dentine caries in permanent molars Follow-up: 2 years	Incidence of carious first molars (40%) 400 per 1000[1]	Incidence of carious first molars (6.3%) 63 per 1000 (38 to 96)	OR 0.12 (0.07 to 0.19) [2]	1259 children randomised & 1066 evaluated after 2 years (6 studies [3,4,5])	⊕⊕⊕○ **moderate**	Benefits of resin-sealant maintained up to at least 48 months of follow-up[6]
	Incidence of carious first molars (70%) 700 per 1000[1]	Incidence of carious first molars (19%) 190 per 1000 (122 to 272)	OR 0.12 (0.07 to 0.19) [2]	1259 children randomised & 1066 evaluated after 2 years (6 studies [3,4,5])	⊕⊕⊕○ **moderate**	Benefits of resin-based sealant maintained up to at least 48 months of follow-up[6]

CI: confidence interval; **OR:** odds ratio

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] The incidence of carious control teeth in the five split-mouth trials included in this comparison ranged from 37% to 69% (studies published between 1976 and 1979). We have shown the effect of sealants at each end of this range. These studies did not give information on the baseline caries prevalence of the children.
The sixth study included in this meta-analysis (parallel group study published in 2012) reported clearly lower incidence of carious first molars than the five split-mouth studies. In sealant group, carious first molars were detected in 9 out of 121 children (7.4%) (11 carious teeth out of 367 sealed teeth) and in placebo group in 21 out of 124 children (17%) (28 carious teeth out of 379 placebo teeth). Caries prevalence: mean baseline dmft level of 3.4.
[2] There was considerable heterogeneity in this estimate (I^2 = 77% P = 0.0007) but all of the trials showed a statistically significant effect favouring sealants.
[3] Six studies at low risk of bias for the four key domains of allocation concealment, incomplete outcome data, selective reporting and baseline comparability of the groups.
[4]All studies recruited children aged 5-10 years. Three studies conducted in areas with fluoridated water, two studies stated water was not fluoridated and the remaining one study did not report whether water supplies were fluoridated.
[5] Five trials were published between 1976 and 1979 and one in 2012. One further parallel group trial from Thailand at unclear risk of bias reporting DFS increment published in 1995 also found a benefit in favour of resin-based sealant (mean difference in DFS increment -0.65, 95% CI -0.83 to -0.47, 276 children evaluated).
[6] The benefit associated with sealant use is maintained at all of the follow-up estimates (up to 9 years) though the number of studies and the number of children available for evaluation reduced markedly over this period (e.g. at 48 to 54 months of follow-up odds ratio 0.21, 95% CI 0.16 to 0.28, two studies at low risk of bias and two studies at high risk of bias, 482 children evaluated; risk ratio 0.24, 95% CI 0.12 to 0.45, one study at unclear risk of bias, 203 children evaluated).

Citations

Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

Not applicable.

1a.8.2. Provide the citation and summary for each piece of evidence.

Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***
1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form 4_NQF_Evidence_6-9.docx

1a.1 <u>For Maintenance of Endorsement:</u> Is there new evidence about the measure since the last update/submission?
Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.
No

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*
<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
Inequalities in oral health status and inadequate use of oral health care services are well documented. Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, L i, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009). Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions.

Evidence-based clinical recommendations recommend that sealants be placed on pits and fissures of children's primary and permanent teeth when it is determined that the tooth, or the patient, is at risk of experiencing caries (Beauchamp et al. 2008). The evidence for sealant effectiveness in permanent molars is stronger than evidence for primary molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013).

The proposed measure, Prevention: Sealants for 6-9 Year-Old Children at Elevated Caries Risk, captures whether children at moderate or high caries risk received a sealant on a permanent first molar tooth. Permanent first molars usually erupt between ages 6 and 7 years. Thus, this measure addresses both the tooth type on which sealants are placed and the timeliness of care provision. The measure Sealants for 6-9 Year-Old Children allows plans and programs to assess whether children at risk for caries are receiving evidence-based prevention and target performance improvement initiatives accordingly.

This measure is a program/plan specific measure that contributes to the Healthy People 2020 Objective OH 12.2 that calls for increasing the percent children aged 6 to 9 years who received dental sealants on one or more of their first permanent molars.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

1b.2. **Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis**. *(<u>This is required for maintenance of endorsement</u>. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Below are the testing data and results that met scientific acceptability criteria for endorsement. Because there were no changes in the data source, level of analysis or setting, additional testing has not been conducted.

Data Sources:

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. Data from calendar years 2010 and 2011 were used for all programs except Florida Medicaid. Full-year data for CY 2011 were not available for Florida Medicaid. Therefore, we report only CY 2010 data for Florida Medicaid.

In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011. [Technically, there were three plans represented in the data because Texas CHIP was served by a single dental plan. Since the program=plan in that case, we included it in the "program" level data.]

Below we provide summary data for each of the five programs and two plans individually.

Programs

Our source data for the testing prior to applying the denominator age criteria of 6-9 years old included children 0-20 years in each program. The number of children ages 0-20 years enrolled at least one month in each program were as follows:

Texas Medicaid, 2011: 3,544,247
Texas Medicaid, 2010: 3,393,963
Texas CHIP, 2011: 842,454
Texas CHIP, 2010: 786,070
Florida CHIP, 2011: 317,146
Florida CHIP, 2010: 315,975
Commercial, 2011: 184,152
Commercial, 2010: 189,968
Florida Medicaid, 2010: 2,068,670

Within these programs, we had claims data available in both years for two dental managed care plans in Florida CHIP. We also report rates for those two plans separately.

Plan 1, 2010: 77,255
Plan 2, 2010: 116,388
Plan 1, 2011: 140,986
Plan 2, 2011: 168,191

The number of children in the age range of 6-9 years specifically were:

Texas Medicaid, 2011: 746,535
Texas Medicaid, 2010: 706,596
Texas CHIP, 2011: 224,908
Texas CHIP, 2010: 210,624
Florida CHIP, 2011: 88,943
Florida CHIP, 2010: 89,897
Commercial, 2011: 36,905
Commercial, 2010: 38,390
Florida Medicaid, 2010: 406,698
Plan 1, 2010: 25,240
Plan 2, 2010: 31,126
Plan 1, 2011: 41,537
Plan 2, 2011: 45,348

Data 1b.2. Performance Scores for Dental Sealants for 6-9 Year-Olds at Elevated Risk

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

Program 1, CY 2011:	23.69%	(	0.2369	,	0.0006	,	0.2357	,	0.2381)
Program 2, CY 2011:	23.01%	(	0.2301	,	0.0017	,	0.2267	,	0.2335)
Program 3, CY 2011:	31.33%	(	0.3133	,	0.0036	,	0.3062	,	0.3204)
Program 4, CY 2011:	22.59%	(	0.2259	,	0.0042	,	0.2176	,	0.2342)
Program 1, CY 2010:	23.38%	(	0.2338	,	0.0007	,	0.2325	,	0.2351)
Program 2, CY 2010:	19.82%	(	0.1982	,	0.0017	,	0.1949	,	0.2015)
Program 3, CY 2010:	30.04%	(	0.3004	,	0.0036	,	0.2933	,	0.3075)
Program 4, CY 2010:	26.68%	(	0.2668	,	0.0043	,	0.2583	,	0.2753)
Program 5, CY 2010:	21.04%	(	0.2104	,	0.0015	,	0.2074	,	0.2134)
Plan 1, CY 2011:	31.43%	(	0.3143	,	0.0054	,	0.3037	,	0.3249)
Plan 2, CY 2011:	30.91%	(	0.3091	,	0.0050	,	0.2993	,	0.3189)
Plan 1, CY 2010:	31.38%	(	0.3138	,	0.0078	,	0.2985	,	0.3291)
Plan 2, CY 2010 :	29.97%	(	0.2997	,	0.0067	,	0.2866	,	0.3128)

The measure rate range of 20% to 30% in CY 2010 (year in which data were available for all five programs) indicates variations in sealant prevalence across programs.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

The measure testing findings are consistent with other data indicating that there are significant variations in the percentage of children who received sealants. Data from the Centers for Medicare and Medicaid Services indicate significant variation among state Medicaid programs, ranging from 6% to 31% of children 6-9 years old, who received a sealant on a permanent molar tooth (Norris 2013; CMS-416 data, FY 2011).

[Complete citations provided in 1c4 and in Evidence Submission Form Template.]

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The same data sources were used as described in 1b.2. The data below summarizes performance data by geographic location and race/ethnicity for CY 2011 (CY 2010 for one program) with the p-values from chi-square tests used to detect whether there were statistically significant differences in performance between groups. Disparities by geographic location were detected for two programs. Statistically significant difference in performance by race and ethnicity also were detected in the two programs for which there were race/ethnicity data. In addition, we also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), Medicaid program type, CHIP dental plan, commercial product line, and preferred language for program communications. We additionally detected disparities by health status, dental plan and Medicaid program type, but data on all of these characteristics were not consistently available for all programs so we are presenting disparities data on those characteristics that were most consistently available and had the greatest standardization

Data1b.4. Disparities in Performance by Geographic Location and Race/Ethnicity

PROGRAM 1	
Overall performance score:	23.69%
Scores by Geographic Location	
Urban:	23.95%
Rural:	21.89%
p-value from Chi-square test:	<.0001
Scores by Race	
Non-Hispanic White:	22.07%
Non-Hispanic Black:	23.08%

Hispanic: 24.31%
p-value from Chi-square test <.0001

PROGRAM 2
Overall performance score: 23.01%
Scores by Geographic Location
Urban: 23.00%
Rural: 23.23%
p-value from Chi-square test: 0.6649
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 3
Overall performance score: 31.33%
Scores by Geographic Location
Urban: 31.29%
Rural: 31.82%
p-value from Chi-square test: 0.7252
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 4
Overall performance score: 22.59%
Scores by Geographic Location
Urban: 22.70%
Rural: 20.60%
p-value from Chi-square test: 0.3436
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 5
Overall performance score: 21.04%
Scores by Geographic Location
Urban: 21.07%
Rural: 19.33%
p-value from Chi-square test: 0.0087
Scores by Race
Non-Hispanic White: 21.24%
Non-Hispanic Black: 19.63%
Hispanic: 21.87%
p-value from Chi-square test <.0001

Note: N/A for race/ethnicity indicates that those programs did not collect race/ethnicity data or had high rates of missing data .

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4**
There is extensive literature documenting disparities in dental service use among children by age, race/ethnicity, and geographic

region, including within vulnerable populations, much of which is summarized in three major national reports on oral health: the Surgeon General's report on Oral Health in America in 2000, the IOM report, Improving Access to Oral Health Care for Vulnerable and Underserved Populations, and the IOM report, Advancing Oral Health in America.

With respect to preventive dental services in general, there are documented disparities. Using data from the National Survey of Children's Health, Edelstein and Chinn (2009) noted disparities in access to preventive dental services by race and income: "Stepwise disparities in access to preventive dental services are evident by race and income in ways that parallel Medical Expenditure Panel Survey findings. White parents report higher use of preventive dental services than do black or Hispanic parents (77%, 66%, and 61%, respectively). Poor parents report less use of services than do low income, middle class, and higher-income parents (58%, 66%, 77%, and 82%, respectively)" (Edelstein & Chinn, 2009, p.418). A recent analysis by Bouchery (2013) of the Medicaid Analytic eXtract files for nine states found variations in the percentage of children receiving a preventive dental visit by age, race and ethnicity, and geographic area. Specifically, relative to the reference group of 9 year olds, the percentage point change in the probability of having a dental preventive services was -27.6 for 3 years old; -8.6 for 6 years, -2.2 for 12 years and -15.4 for 15 years (all significant at $p<0.0001$); relative to the reference group of white, non-Hispanic, the percentage point change was -1.8 for black non-Hispanic and 7.8 for Hispanic ($p<0.0001$ for both); relative to the reference group of small metro area, the percentage point change was 5.9 for large metro area ($p<0.0001$).

In addition, there are documented disparities in dental sealant receipt specifically. For example, using data from the National Health and Nutrition Examination Survey, researchers at the National Center for Health Statistics identified variations in dental sealant prevalence among children by age, race, ethnicity, and poverty level (Dye, Li, and Thorton-Evans 2012). Specifically: "Dental sealant prevalence was lower among children [6-9 years] living at or below 100% of the federal poverty level (26%) compared with children living above the poverty level (34%). A similar pattern was found among adolescents aged 13–15, but the difference was not statistically significant. Dental sealant prevalence was significantly lower for non-Hispanic black adolescents (32%) compared with non-Hispanic white adolescents (56%), among those aged 13–15" (Dye, Li, and Thorton-Evans 2012, p. 2).

Sources

Bouchery, E. 2013. "Utilization of Dental Services among Medicaid-Enrolled Children." Medicare & Medicaid Research Review. 3(3) E1-16. Available at: https://www.cms.gov/mmrr/Downloads/MMRR2013_003_03_b04.pdf.

Dietrich, T., C. Culler, R. Garcia, and M. M. Henshaw. 2008. Racial and ethnic disparities in children's oral health: The National Survey of Children's Health. Journal of the American Dental Association 139(11):1507-1517.

Dye BA, Li X, Thorton-Evans G. Oral health disparities as determined by selected healthy people 2020 oral health objectives for the United States, 2009-2010. NCHS Data Brief 2012(104):1-8.U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research.

Edelstein, B. L. and C. H. Chinn. 2009. "Update on Disparities in Oral Health and Access to Dental Care for America´s Children." Acad Pediatr 9(6): 415-9.

Institute of Medicine (U.S.). Committee on an Oral Health Initiative. Advancing oral health in America. Washington, D.C.: National Academies Press; 2011.

Institute of Medicine and National Research Council. Improving access to oral health care for vulnerable and underserved populations. Washington, D.C.: National Academies Press; 2011.

Kenney, G. M., J. R. McFeeters, and J. Y. Yee. 2005. Preventive dental care and unmet dental needs among low-income children. American Journal of Public Health 95(8):1360-1366.

Lewis, C., W. Mouradian, R. Slayton, and A. Williams. 2007. Dental insurance and its impact on preventative dental care visits for U.S. children. Journal of the American Dental Association 138(3):369-380.

Oral Health in America: a report of the Surgeon General. Rockville, Md.: U.S. Public Health Service, Dept. of Health and Human Services; 2000.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Dental

De.6. **Non-Condition Specific***(check all the areas that apply):*
Access to Care, Disparities Sensitive, Health and Functional Status : Change, Health and Functional Status : Total Health, Primary Prevention

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, Populations at Risk

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.ada.org/~/media/ADA/Science%20and%20Research/Files/DQA_2018_Dental_Services_Sealants_6-9_Years.pdf?la=en

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
1. No changes to the measure specifications

2. Measure specification website updated to be more user friendly

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Unduplicated number of enrolled children age 6-9 years at "elevated" risk (i.e., "moderate" or "high") who received a sealant on a permanent first molar tooth as a dental service.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see section S14

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Unduplicated number of enrolled children age 6-9 years who are at "elevated" risk (i.e., "moderate" or "high")

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see section S14

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Medicaid/ CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded.

There are no other exclusions.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
There are no other exclusions than those described above.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
There are no stratifications for this measure.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Sealants for 6 – 9 year olds - Calculation for Children at Elevated Caries Risk

1. Use administrative enrollment and claims data for a single year. When using claims data to determine service receipt, include both paid and unpaid claims (including pending, suspended, and denied claims).

2. Check if the enrollee meets age criteria at the last day of the reporting year:

a. If child is >= 6 and <= 9, then proceed to next step.
b. If age criterion is not met or there are missing or invalid field codes (e.g., date of birth), then STOP processing. This enrollee does not get counted.

3. Check if subject is continuously enrolled for at least 180 days during the reporting year:
a. If subject meets continuous enrollment criterion, then proceed to next step.
b. If subject does not meet enrollment criterion, then STOP processing. This enrollee does not get counted.

YOU NOW HAVE THE COUNT OF THOSE WHO MEET THE AGE AND ENROLLMENT CRITERIA

4. Check if subject is at "elevated risk":
i. the subject has a CDT Code among those in Table 1 in the reporting year,
OR
ii. the subject has a CDT Code among those in Table 1 in any of the three years prior to the reporting year, (NOTE: The subject does not need to be enrolled in any of the prior three years for the denominator enrollment criteria; this is a "look back" for enrollees who do have claims experience in any of the prior three years.)
OR
iii. the subject has a visit with a CDT code = (D0602 or D0603) in the reporting year.
b. If the subject does not meet any of the above criteria for elevated risk, then STOP processing. This enrollee will not be included in the measure denominator.

YOU NOW HAVE THE DENOMINATOR (DEN): Enrollees who are at "elevated risk"

5. Check if subject received a sealant as a dental service:
a. If [CDT CODE] = D1351 and;
b. If [RENDERING PROVIDER TAXONOMY] code = any of the NUCC maintained Provider Taxonomy Codes in Table 2 below, then proceed to next step.
c. If both a AND b are not met, then the service was not a "dental service"; STOP processing. This enrollee is already included in the denominator but will not be included in the numerator.

Note: In this step, all claims with missing or invalid CDT CODE, missing or invalid NUCC maintained Provider Taxonomy Codes, or NUCC maintained Provider Taxonomy Codes that do not appear in Table 2 should not be included in the numerator.

6. Check if sealant was placed on a permanent first molar:
a. If [TOOTH-NUMBER] = 3, 14, 19 or 30 then include in numerator; STOP processing.
b. If not, then service was not provided for the first permanent molar; STOP processing. This enrollee is already included in the denominator but will not be included in the numerator.

YOU NOW HAVE NUMERATOR (NUM) COUNT: Enrollees at "elevated risk" who received sealants on a permanent first molar as a dental service

7. Report
a. Unduplicated number of enrollees in numerator
b. Unduplicated number of enrollees in denominator
c. Measure rate (NUM/DEN)

Table 1: CDT Codes to identify "elevated risk"

D2140	D2394	D2630	D2720	D2791	D3120
D2150	D2410	D2642	D2721	D2792	D3220
D2160	D2420	D2643	D2722	D2794	D3221
D2161	D2430	D2644	D2740	D2799	D3222
D2330	D2510	D2650	D2750	D2930	D3230
D2331	D2520	D2651	D2751	D2931	D3240
D2332	D2530	D2652	D2752	D2932	D3310
D2335	D2542	D2662	D2780	D2933	D3320
D2390	D2543	D2663	D2781	D2934	D3330
D2391	D2544	D2664	D2782	D2940	D2941

D2392 D2610 D2710 D2783 D2950 D1354
D2393 D2620 D2712 D2790 D3110

Table 2: NUCC maintained Provider Taxonomy Codes classified as "Dental Service"*

122300000X	1223P0106X	1223X0008X	261QF0400X
1223D0001X	1223P0221X	1223X0400X	261QR1300X
1223D0004X	1223P0300X	124Q00000X+	125Q00000X
1223E0200X	1223P0700X	125J00000X	
1223G0001X	1223S0112X	125K00000X	

*Services provided by County Health Department dental clinics may also be included as "dental" services.
+Only dental hygienists who provide services under the supervision of a dentist should be classified as "dental" services. Services provided by independently practicing dental hygienists should be classified as "oral health" services and are not applicable for this measure.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
Not applicable.

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
Not applicable.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
Not applicable.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable.

2. Validity – See attached Measure Testing Submission Form
5_Testing_6-9.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Prevention: Dental Sealants for 6-9 Year-Old Children at Elevated Caries Risk
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.

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] 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] 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]

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]

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; [14,15] 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] **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
☒ 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:	☐ other:

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 <u>measured entities</u> 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*)

<u>Level of Analysis: Program, 5 Measured Entities</u>

1. Texas Medicaid
   A. 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
   A. 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
   A. 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
   A. 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
   A. Size: # Members 0-20 years, CY 2010: 2,068,670;
   B. 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.

<u>Level of Analysis: Plan, 2 Measured Entities</u>

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

2) FL CHIP – Plan 2
   A. 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 <u>patients</u> 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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2011					
	Program 1	Program 2	Program 3	Program 4	Plan 1	Plan 2
Total Number Patients	3,544,247	842,454	317,146	184,152	140,986	168,191
Age Group Distribution						
Age <1 years	7.05%	0.11%	N/A	1.54%	N/A	N/A
Age 1-2 years	14.32%	5.34%	N/A	5.75%	N/A	N/A
Age 3-5 years	19.46%	11.70%	3.81%	12.68%	4.12%	3.60%
Age 6-7 years	11.21%	12.30%	13.05%	9.57%	13.71%	12.55%
Age 8-9 years	9.85%	14.40%	15.00%	10.18%	15.76%	14.41%
Age 10-11 years	9.03%	14.03%	15.71%	10.55%	16.27%	15.25%
Age 12-14 years	11.63%	19.57%	23.73%	16.09%	23.06%	24.31%
Age 15-18 years	13.19%	22.54%	28.70%	22.13%	27.08%	29.88%
Age 19-20 years	4.27%	N/A	N/A	11.50%	N/A	N/A
Geographic Location						
Urban	83.63%	84.33%	92.94%	95.95%	93.01%	92.91%
Rural	15.15%	14.61%	5.02%	3.86%	4.83%	5.15%
Missing	1.22%	1.06%	2.04%	0.19%	2.16%	1.94%
Race and Ethnicity						
Non-Hispanic White	17.36%	N/A	N/A	N/A	N/A	N/A
Non-Hispanic Black	15.08%	N/A	N/A	N/A	N/A	N/A
Hispanic	58.07%	N/A	N/A	N/A	N/A	N/A
Other & Unknown	9.49%	N/A	N/A	N/A	N/A	N/A

Table 1.6B, Patient Characteristics, 6-9 Years Old (Age Range Targeted by Measure), 2011

	Descriptive Characteristics of Individuals 6-9 Years Enrolled at Least One Month, CY 2011					
	Program 1	Program 2	Program 3	Program 4	Plan 1	Plan 2
Total Number Patients	746,535	224,908	88,943	36,905	41,537	45,348
Age Group Distribution						
Age 6-7 years	53.23%	46.05%	46.53%	48.49%	46.52%	46.55%
Age 8-9 years	46.77%	53.95%	53.47%	51.51%	53.48%	53.45%
Geographic Location						
Urban	84.16%	84.46%	93.32%	96.19%	93.41%	93.29%
Rural	15.00%	14.54%	5.00%	3.58%	4.84%	5.11%
Missing	0.84%	1.00%	1.68%	0.23%	1.75%	1.61%
Race and Ethnicity						
Non-Hispanic White	16.77%	N/A	N/A	N/A	N/A	N/A
Non-Hispanic Black	14.90%	N/A	N/A	N/A	N/A	N/A
Hispanic	62.04%	N/A	N/A	N/A	N/A	N/A
Other & Unknown	6.29%	N/A	N/A	N/A	N/A	N/A

Table 1.6C, Patient Characteristics, 0-20 Years Old, 2010

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2010						
	Program 1	Program 2	Program 3	Program 4	Program 5	Plan 1	Plan 2
Total Number Patients	3,393,963	786,070	315,975	189,968	2,068,670	77,255	116,388
Age Group Distribution							
Age <1 years	7.35%	0.15%	N/A	1.45%	6.05%	N/A	N/A
Age 1-2 years	15.16%	5.37%	N/A	5.67%	14.23%	N/A	N/A
Age 3-5 years	19.48%	11.69%	3.64%	12.73%	19.26%	5.72%	4.22%
Age 6-7 years	11.12%	12.19%	13.32%	9.69%	10.47%	15.68%	12.54%
Age 8-9 years	9.70%	14.61%	15.14%	10.24%	9.19%	16.99%	14.21%
Age 10-11 years	8.75%	14.04%	15.84%	10.60%	8.74%	16.41%	15.18%
Age 12-14 years	11.23%	19.49%	23.70%	16.20%	11.87%	21.40%	24.05%
Age 15-18 years	12.99%	22.47%	28.37%	22.12%	14.73%	23.79%	29.81%
Age 19-20 years	4.22%	N/A	N/A	11.31%	5.47%	N/A	N/A
Geographic Location							
Urban	83.20%	84.46%	92.08%	96.70%	91.47%	92.10%	92.11%
Rural	15.56%	14.45%	5.07%	3.17%	7.30%	5.00%	5.19%
Missing	1.24%	1.08%	2.85%	0.13%	1.23%	2.89%	2.70%
Race and Ethnicity							
Non-Hispanic White	18.21%	N/A	N/A	N/A	29.89%	N/A	N/A
Non-Hispanic Black	15.45%	N/A	N/A	N/A	29.39%	N/A	N/A
Hispanic	59.42%	N/A	N/A	N/A	29.65%	N/A	N/A
Other & Unknown	6.92%	N/A	N/A	N/A	11.06%	N/A	N/A

Table 1.6D, Patient Characteristics, 6-9 Years Old (Age Range Targeted by Measure), 2010

	Descriptive Characteristics of Individuals 6-9 Years Enrolled at Least One Month, CY 2010						
	Program 1	Program 2	Program 3	Program 4	Program 5	Plan 1	Plan 2
Total Number Patients	706,596	210,624	89,897	36,905	406,698	25,240	31,126
Age Group Distribution							
Age 6-7 years	53.39%	45.48%	46.80%	48.49%	53.23%	47.98%	46.88%
Age 8-9 years	46.61%	54.52%	53.20%	51.51%	46.77%	52.02%	53.12%
Geographic Location							
Urban	83.80%	84.66%	92.92%	96.19%	91.39%	93.10%	92.97%
Rural	15.35%	14.28%	5.08%	3.58%	7.35%	4.86%	5.12%
Missing	0.85%	1.06%	1.99%	0.23%	1.26%	2.04%	1.91%
Race and Ethnicity							
Non-Hispanic White	17.00%	N/A	N/A	N/A	29.57%	N/A	N/A
Non-Hispanic Black	14.85%	N/A	N/A	N/A	29.19%	N/A	N/A
Hispanic	62.25%	N/A	N/A	N/A	30.95%	N/A	N/A
Other & Unknown	5.89%	N/A	N/A	N/A	10.30%	N/A	N/A

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) the evidence regarding the effectiveness of sealants in caries prevention, (2) an environmental scan, 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. 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
Dental Sealants for 6-9 Year-Old Children at Elevated Caries Risk measures the percentage of children ages 6-9 years at moderate to high risk for dental caries who received a sealant on a permanent first molar tooth during the reporting year. 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) 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

# Requested	# Received	%Received
600	414	69%

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 sealant procedure and tooth number 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 - Sealant and Tooth Number Codes

Data Extraction. For the second component of validation, assessing whether the specific preventive service of sealant placement and associated tooth type coding are accurately captured by claims data, chart abstraction forms were developed by the research team. 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 sealants applied to the teeth during the visit. If there was evidence of sealant placement, the RHITs then recorded whether sealants were applied to the child's permanent first molar, permanent second molar, and/or "other" tooth type. If there was no indication of the tooth to which the sealant was applied, the tooth number field was coded as "indeterminate." 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 code D1351 (sealants); and, when D1351 was present, recording the associated tooth number (or noted as missing). Permanent first molars were identified in the claims data as tooth numbers

3, 14, 19, and 30; permanent second molars were identified as tooth numbers 2, 15, 18, and 31. 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 - RELEVANCE OF TOOTH TYPE

Evidence-based recommendations advise that sealants be placed on pits and fissures of children's primary and permanent teeth when the tooth, or patient, is at caries risk, with stronger evidence for effectiveness in permanent molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013). Thus, we also sought to evaluate how well the specifications addressed both the tooth type on which sealants are placed and the timeliness of care provision. The research team ran frequency distributions of sealant placement by tooth number and age range for three programs. Specifically, the percentage of children with (1) any sealants (regardless of tooth type), (2) sealants on permanent first molars, and (3) sealants on permanent second molars was assessed by age for children enrolled at least one month in the program.

Citations

Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268.

Centers for Disease Control and Prevention. 2013. Dental Sealants. Available at: http://www.cdc.gov/OralHealth/publications/faqs/sealants.htm. Accessed January 20, 2014.

U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research. Oral health in America : a report of the Surgeon General. Rockville, Md.: U.S. Public Health Service, Dept. of Health and Human Services; 2000.

4. 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 preventive care 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. 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 preventive care 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

	Percentage of All Enrolled Members Included in Different Denominator Definitions				
	Program 1	Program 2	Program 3	Program 4	Program 5
At least 1 month	100%	100%	100%	100%	100%
At least 3 months	95%	85%	84%	93%	94%
At least 6 months	83%	63%	65%	81%	81%
11-12 months	64%	33%	42%	63%	59%

5. ADDITIONAL VALIDITY TESTING - IDENTIFYING ELEVATED RISK WITH CLAIMS DATA

Evidence-based guidelines indicate that sealants are most effective for children at higher risk for caries (see Measure Evidence Form). Thus, inclusion in the denominator is limited to children identified as being at moderate to high risk for caries. Administrative claims data for dental claims typically do not include diagnostic codes. Procedure codes for risk assessment that identify moderate and high risk were included in the measure logic. However, because these are newer codes, additional logic was included to identify children with recent history of restorations, which are indicative of caries. A systematic review found that prior caries experience to be an important predictor of future risk (Zero D, Fontana M, Lennon AM. 2001. Clinical applications and outcomes of using indicators of risk in caries management. J Dent Educ. 2001 Oct;65(10):1126-32.) Expert consensus and validation through chart reviews was done to finalize the procedure codes (indicated in the measure specifications) used to identify elevated risk. The test data results reported in this application demonstrate that it is feasible to use these validated codes to identify children at elevated risk who should receive preventive services.

6. 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

Number of Procedure Codes	Record and Procedure Code on Claim Correlate	Record Did Not Correlate with Procedure Code on Claim	Unable to Determine Correlation
1,135	94.04%	4.22%	1.75%

B. Critical Data Element Validation - Sealant and Tooth Number Codes

To assess whether the specific preventive service of dental sealants and associated tooth type 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 for sealants and tooth number. Agreement (concordance) for sealant placement was 95%. Sensitivity of sealant placements was moderately high (77.8%) and specificity was very high (98.8%). Similar findings were obtained for first molars. The positive predictive and negative predictive values were both high (>93%) for sealant placement with a lower negative predictive value for the specific tooth type. 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 for sealants was also very high at 0.8205 indicating "almost perfect" agreement. For dates of service in which there was agreement with the administrative data that sealants had been applied (n=84), we then assessed whether there was agreement on tooth type using the following categories: permanent first molar, permanent second molar, and other teeth. We report here on the findings for permanent first molar which is the focus of the proposed measure (we had similar findings for second molars). Overall, the simple agreement percentage was 84% for permanent first molars. The corresponding kappa statistic value was 0.691, indicating "substantial" agreement.

Table 2b2.3-2 Agreement between Record and Administrative Data for Specific Services

	Concordance	Prevalence	Sensitivity	Specificity	PPV	NPV	Kappa
Sealants Applied	95.22%	0.172	0.778	0.988	0.933	0.955	0.820
Dates of service: 613			(0.686-0.850)	(0.974-0.995)	(0.855-0.973)	(0.933-0.971)	(0.758-0.882)
# indeterminate: 4							
First Molar (if sealant)	84.34%	0.627	0.750	1.000	1.000	0.705	0.691
Dates of service: 613			(0.608-0.855)	(0.863-1.000)	(0.888-1.000)	(0.546-0.828)	(0.545-0.838)
# indeterminate: 1							

95% confidence intervals indicated in parentheses

Our findings are similar 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. For sealants,

they found lower sensitivity (73%), higher specificity (100%) and similar kappa value (0.84) of billing data compared to direct observation. (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

Sealants on a Permanent Molar Tooth was identified through the Delphi rating process as a high-scoring measure concept with a mean importance score of 7, mean feasibility score of 8, and mean validity score of 7, 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.] Thus, the measure has face validity. However, gaps were identified with existing measures, including not associating tooth type and age range, lack of clear specifications, and lack of standardization. The proposed measure overcomes these limitations.

Content Validity. In addition, the measure also demonstrates **content validity** – the extent to which the measure specifications reflect the intended domain of care. This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (sealants) as well as the specific tooth type for which the evidence is the strongest (permanent molar) and the timing of sealant placement to maximize effectiveness (shortly after eruption – 6-9 years of age for permanent first molars). Please see the Measure Evidence Form for more details.

3. ADDITIONAL VALIDITY TESTING - RELEVANCE OF TOOTH NUMBER

Analysis of sealant placement by tooth type and age range validated the importance of including specific teeth numbers in the measure specifications to identify permanent first molars and permanent second molars and associating those tooth numbers with the corresponding appropriate age ranges (6-9 years and 10-14 years, respectively) in order to have reliable indicators of whether children are getting recommended and timely prevention. Table 2b2.3-3 indicates the percentage of children in each of three programs who had (1) a sealant placed on any tooth, (2) a sealant placed on a permanent first molar, and (3) a sealant placed on a permanent second molar; the same child could be included in more than one category. In programs 3 and 4, the percentage of children ages 6-9 years with sealants on permanent first molars is very close to the percentage of children with sealants on any tooth, suggesting that most children ages 6-9 years in these two programs who received sealants received them for permanent first molars. However, in Program 1 there were substantial differences between the percentage of children with a sealant on any tooth compared to the percentage of children with a sealant on a permanent first molar. For example, 25% of children received a sealant, but only 14% received a sealant specifically on a permanent first molar. The differences reflect differences in benefit coverage between the programs; Program 1did not condition reimbursement for sealants on tooth type. These results indicate that children ages 6-9 years may have teeth other than permanent first molars (e.g., premolars or primary teeth) sealed that would get captured in the numerator and inflate the measure score if teeth numbers are not included, resulting in misleading comparisons of performance between programs. Thus, the research team concluded that the incorporation of teeth numbers in the DQA specifications is a significant and important improvement over existing sealant measures that have lacked this specificity.

Table 2b2.3-3 Sealant Placement by Age and Tooth Type

	Program 1			Program 3			Program 4		
Age (years)	% with Any Sealants (Any Tooth)	% with Sealant on Permanent 1st Molars	% with Sealant on Permanent 2nd Molars	% with Any Sealants (Any Tooth)	% with Sealant on Permanent 1st Molars	% with Sealant on Permanent 2nd Molars	% with Any Sealants (Any Tooth)	% with Sealant on Permanent 1st Molars	% with Sealant on Permanent 2nd Molars
6	25.02%	13.73%	0.04%	6.42%	6.32%	0.04%	8.21%	7.58%	0.01%
7	34.44%	26.20%	0.06%	15.03%	14.95%	0.09%	21.21%	20.92%	0.09%
8	31.02%	21.56%	0.08%	15.52%	15.49%	0.15%	18.85%	18.70%	0.12%
9	29.80%	14.00%	0.28%	12.45%	12.34%	0.18%	11.35%	11.06%	0.19%
10	35.36%	9.91%	1.87%	10.36%	9.90%	0.88%	7.63%	6.77%	0.74%
11	40.45%	7.42%	6.92%	10.18%	8.78%	3.07%	7.70%	4.92%	3.18%
12	40.96%	5.36%	12.76%	10.46%	7.67%	6.29%	11.99%	4.57%	9.05%
13	36.20%	3.73%	14.40%	10.40%	6.89%	8.27%	14.94%	4.04%	13.34%
14	29.85%	2.82%	11.64%	9.07%	5.93%	8.08%	12.44%	3.32%	11.51%

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?*)
As noted above, the overall agreement between the administrative claims data and dental record data was high based on both simple agreement and using the more conservative Kappa statistic. Although the agreement for the specific tooth type was not as strong as for sealant application in general, it was still "substantial," and we believe that data concordance will improve with increasing accountability as is often the case when new performance measures are implemented. Overall, we interpret these findings as evidence that validates the accuracy of administrative claims data for performance measurement purposes. These empirical findings, combined with our face validity assessments of the measure score, lead us to conclude that both the data elements and the measure score represent valid measures of sealant placement prevalence among 6-9 year olds. In addition, our testing indicated that the incorporation of tooth number as part of the measure specifications was important for ensuring that the measure captures sealant placement on the tooth type (permanent first molars) for which there is the strongest evidence of effectiveness among this age group.

2b3. EXCLUSIONS ANALYSIS
NA X☐ no exclusions — *skip to section 2b4*

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.
Not applicable.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with risk factors**
☐ **Stratification by risk categories**
☐ **Other,**

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*)
Not applicable.

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

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 had a sealant, the dichotomous outcome of had/did not have a sealant 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/Plan, Year, Measure Score as % (Measure Score, SD, Lower 95% CI, Upper 95% CI)

Program 1, CY 2011:	23.69%	(0.2369 ,	0.0006 ,	0.2357 ,	0.2381)
Program 2, CY 2011:	23.01%	(0.2301 ,	0.0017 ,	0.2267 ,	0.2335)
Program 3, CY 2011:	31.33%	(0.3133 ,	0.0036 ,	0.3062 ,	0.3204)
Program 4, CY 2011:	22.59%	(0.2259 ,	0.0042 ,	0.2176 ,	0.2342)
Program 1, CY 2010:	23.38%	(0.2338 ,	0.0007 ,	0.2325 ,	0.2351)
Program 2, CY 2010:	19.82%	(0.1982 ,	0.0017 ,	0.1949 ,	0.2015)
Program 3, CY 2010:	30.04%	(0.3004 ,	0.0036 ,	0.2933 ,	0.3075)
Program 4, CY 2010:	26.68%	(0.2668 ,	0.0043 ,	0.2583 ,	0.2753)
Program 5, CY 2010:	21.04%	(0.2104 ,	0.0015 ,	0.2074 ,	0.2134)
Plan 1, CY 2011:	31.43%	(0.3143 ,	0.0054 ,	0.3037 ,	0.3249)
Plan 2, CY 2011:	30.91%	(0.3091 ,	0.0050 ,	0.2993 ,	0.3189)
Plan 1, CY 2010:	31.38%	(0.3138 ,	0.0078 ,	0.2985 ,	0.3291)
Plan 2, CY 2010 :	29.97%	(0.2997 ,	0.0067 ,	0.2866 ,	0.3128)

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 in both years (Table 2b5.2).

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

	Chi-Square Value	*p* - value
Program Results, 2011	548.60	<0.0001
Program Results, 2010	1049.18	<0.0001
Plan Results, 2011	0.50	0.4795
Plan Results, 2010	1.88	0.1703

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 the program level with

performance scores ranging by approximately 10 percentage points. We did not detect statistically significant differences between the two plans within FL CHIP for this measure. Performance between the two plans were similar on this measure with a 1/2 of one percentage point difference in 2010 (31.43% versus 30.91%) and a 1.41 percentage point difference in 2011 (31.39% versus 29.97%). We do not believe that this signifies the inability of the measure to detect differences in performance between plans; rather, the two plans we tested performed similarly on the measure. Presumably, testing does not require that all comparisons evaluated demonstrate statistically significant differences; rather, testing should demonstrate that where meaningful differences exist, they can be detected. However, we can also look to Program 2 for further comparisons at the plan level because Program 2 was served by a single dental plan so the program measure score also represents a plan-level score. Differences between the Program 2 measure scores (which also represents a single dental plan) are significantly different from those for Plan 1 and Plan 2 as can be seen by comparing the confidence intervals in Table 1b.2. Collectively, these findings are consistent with evidence reported elsewhere in this application documenting disparities in sealant receipt among children. Thus, this measure informs performance improvement efforts by allowing plans and programs to identify and monitor performance gaps and disparities in performance 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.

3. Feasibility
Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
3a. Byproduct of Care Processes For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
This measure is specified for reporting at the program and plan level and there are currently no efforts to develop an eMeasure (eCQM) at the same reporting level.

Our understanding is that the Feasibility Score Card is only for eMeasures; consequently, we have not submitted this. Feasibility criteria were met during the initial endorsement review.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
This measure relies on standard data elements in administrative claims data (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes. A key advantage of using administrative claims data is that the time and cost of data collection for performance measurement purposes are relatively low because these data are already collected for other purposes.

Initial feasibility assessments were conducted using the RAND-UCLA modified Delphi process to rate the measure concepts with feasibility as one component of the assessment. On a 1-9 point scale, this measure concept was rated as an 8 or "definitely feasible" by the expert panel. During the empirical testing phase, our testing found that all of the critical data elements except one had missing/invalid data of <1% (Data 3c.1.), meeting or exceeding the guidance from the Centers for Medicare and Medicaid Services regarding acceptable error rates. The exception was tooth number associated with sealant procedure codes. Missing/invalid data rates ranged from 0.15% to 15%, with most programs having missing/invalid rates <5%. We do not view the higher rates among a subset of the programs as a threat to feasibility, however. The high compliance by the majority of programs indicates that it is feasible to obtain missing and invalid rates of <1%. The Centers for Medicare and Medicaid Services already requires state Medicaid programs to report sealants placed on permanent molars among enrolled children, which requires data on tooth number, and tooth number also is typically required for reimbursement. During measure development and testing, the measure specifications were made available through a publicly accessible website for public comment with additional broad email dissemination to a wide range of stakeholders. No concerns regarding feasibility of collecting any of the data elements were raised during this process.

Citation: Centers for Medicare & Medicaid Services. Medicaid and CHIP Statistical Information System (MSIS) File Specifications and

Data Dictionary. 2010; http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MSIS/downloads/msisdd2010.pdf. Accessed August 10, 2013.

Data 3c.1 Percentage of Missing and Invalid Values for Critical Data Elements

PROGRAM 1
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Tooth number: 6.18%
Date of Service: 0.01%
Rendering Provider ID: 0.28%

PROGRAM 2
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Tooth number: 15.31%
Date of Service: 0.00%
Rendering Provider ID: 0.00%

PROGRAM 3
Member ID: 0.27%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.28%
Tooth number: 0.18%
Date of Service: 0.00%
Rendering Provider ID: 0.18%

PROGRAM 4
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.01%
Tooth number: 2.47%
Date of Service: 0.00%
Rendering Provider ID: 0.61%

PROGRAM 5
Member ID: 0.43%
Date of Birth: 0.02%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Tooth number: 0.15%
Date of Service: 0.00%
Rendering Provider ID: 0.67%

Endorsement Maintenance Update:
This measure is included in the CHIPRA Core Measures Program. Some Medicaid programs noted that they do not receive complete data on tooth number from their contracted plans, which is a required data element for this measure. As a result, the affected programs must get these data from their contracted plans. Because tooth number is required for reimbursement, these data are readily accessible for plan level reporting. Despite initial concerns about this data element, 25 states reported this measure in FFY 2015, and 34 reported in FFY 2016.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

This measure is intended to be transparent and available for widespread adoption. As such, it was purposefully designed to avoid using software or other proprietary materials that would require licensing fees. The measure specifications, including a companion User Guide, are accessible through a website and can be used free of charge for non-commercial purposes. The main requirement of users is to ensure the quality of their source data and expertise to program the measures within their information systems, following the clear and detailed specifications. Technical assistance is available to users.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting CMS CHIPRA core set https://www.medicaid.gov/medicaid/quality-of-care/downloads/2017-child-core-set.pdf Texas Health and Human Services Commission: Texas Medicaid/CHIP https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf Payment Program Texas Health and Human Services Commission: Texas Medicaid/CHIP https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf Quality Improvement (external benchmarking to organizations) Covered California http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf CMS CHIPRA core set https://www.medicaid.gov/medicaid/quality-of-care/downloads/2017-child-core-set.pdf Michigan Healthy Kids Dental https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids Quality Improvement (Internal to the specific organization) State Medicaid Agencies http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

4a1.1 **For each CURRENT use, checked above (update for maintenance of endorsement), provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

1. Program and Sponsor: Texas Health and Human Services Commission - Texas Medicaid and CHIP

https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf

Purpose: Payment Program/Public Reporting

This measure has been adopted by the Texas Health and Human Services Commission as part of the Texas CHIP and Medicaid Dental Services Pay-for-Quality (P4Q) program. [Texas HHSC Uniform Managed Care Manual, Chapters 6.2.15. Effective Date 09/01/2017, Version 2.0].

This measure was also present in earlier iterations of the Texas Medicaid and CHIP quality programs since initial endorsement. We are referencing current use for this update.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies to the state of Texas CHIP and Medicaid programs (statewide application). There are two dental plans (i.e., the accountable entities) that serve Texas CHIP and Medicaid. In June 2017, there were 3,359,770 children enrolled in Texas Medicaid and CHIP (https://hhs.texas.gov/about-hhs/records-statistics/data-statistics/healthcare-statistics).

Level of Measurement and Setting: The measure is implemented at the plan and program levels within the Texas Medicaid and CHIP programs.

2. Covered California, the California Health Benefit Exchange

http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf
http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-QDP-Issuer-Contract-and-Attachments.pdf

Purpose: Quality Improvement

This measure is included in the Covered California Qualified Health Plan Issuer Contract for 2017-019 For the Individual Market and the Covered California Qualified Dental Plan Issuer Contract for 2017-2019. The measure is to be reported annually.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies statewide. In March 2017 there were 85,000 enrollees 0-18 years old in CC health plans (which may offer dental benefits and would therefore report on the dental quality measures). There were 5,100 children enrolled specifically in Qualified Dental Plans. (http://hbex.coveredca.com/data-research/)

Level of Measurement and Setting. The measure is implemented at the plan level with the Covered California program.

3. Centers for Medicare and Medicaid Services, Core Set of Children's Health Care Quality Measures for Medicaid and CHIP (CMS CHIPRA Core Set)

https://www.medicaid.gov/medicaid/quality-of-care/downloads/2017-child-core-set.pdf

Purpose: Quality Improvement/Public Reporting

This measure was included in the CHIPRA Core Set, with reporting starting in FFY 2015. In the first year of reporting, 25 states reported this measure (https://www.medicaid.gov/medicaid/quality-of-care/downloads/performance-measurement/2016-child-chart-pack.pdf). In the second year of reporting (FFY 2016), 34 states reported this measure (https://data.medicaid.gov/Quality/2016-Child-Health-Care-Quality-Measures/wnw8-atzy).

Geographic Area and Number/Percentage of Accountable Entities and Patients:
34 states are currently reporting this measure. Information is not provided on the number of accountable entities and patients.

4. State Medicaid Agencies

http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

(Note: To access the data, a public user account must be created. We can help facilitate access to the data if needed.)

Purpose: Quality Improvement

The Medicaid|Medicare|CHIP Services Dental Association conducts an annual survey of state Medicaid programs and collects data specifically on which programs report Dental Quality Alliance measures.

In its 2015 profile (the most recent available), 13 states reported that they currently use this measure in their Medicaid and/or CHIP programs.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The 13 states are: Alabama, Colorado, Connecticut, Florida, Idaho, Illinois, Nevada, Oklahoma, Rhode Island, South Carolina, Tennessee, Virginia,and West Virginia. Data are not provided on the number of accountable entities included.

5. Michigan Healthy Kids Dental
https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids

Note: Select Schedule A Work Statement link under File Attachments

Purpose: Quality Improvement

The Michigan Healthy Kids Dental Program has included this measure in the set of measures included in its Performance Monitoring Standards, which is currently included in the Request for Proposals and will be included in the contracts between the contracted dental plans and the State of Michigan.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The Healthy Kids Dental Program covers children enrolled in Michigan's Medicaid program statewide. The state intends to award two contracts. There are approximately 955,000 enrollees served by the Healthy Kids Dental Program.

Additional Information:
This measure was one of ten performance measures approved by the Dental Quality Alliance (DQA) that focused on Dental Caries Prevention and Disease Management among children. 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)

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not applicable.

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
This measure is part of the CMS CHIPRA core set for public reporting by all state CHIP programs. In FFY 2016, 34 states reported on this measure. States also report using this measure in the annual survey conducted by the Medicaid|Medicare|CHIP Services Dental Association. The measure is part of measure set included in the Request for Proposals (RFP) released by the Michigan Healthy Kids Dental Program. This measure is included in the Pay-For-Quality program and public reporting in the Texas Medicaid and CHIP programs. Additionally, this measure is a requirement for the Qualified Dental Plans to report to the Covered California, the state-based marketplace in California.

The DQA provides technical assistance to these and other users of DQA measures through webinars, resource document development, and one-on-one staff support. The DQA has an Implementation Committee dedicated to developing implementation and improvement resources.

In order to ensure transparency, incorporate learnings from implementation, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting. Implementation assistance has mostly focused on addressing questions related to how to use the measures in the context of broader quality improvement and clarifying questions related to the specifications.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

The DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.

In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.
DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
A dental benefits administrator (DBA) has suggested that the DQA consider adding patient exclusions to the measure. The DQA considered exclusions previously during initial measure development and during annual reviews. Exclusions were not incorporated due to concerns about the introduction of biased measurement, increasing measurement complexity, and adversely affecting implementation feasibility. However, the DQA continues to monitor this issue and will revisit it during the 2018 annual review. The DQA has invited the DBA to present its suggestion with supporting data to the DQA. The DQA has also invited other DBAs and Medicaid program administrators to provide input. All of this stakeholder feedback will be incorporated into the next annual review.

4a2.2.3. **Summarize the feedback obtained from other users**
No other significant issues have been raised by other users.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
The DQA considered exclusions during initial measure development and during annual reviews. Exclusions were not incorporated due to concerns about the introduction of biased measurement, increasing measurement complexity, and adversely affecting implementation feasibility. However, the DQA continues to monitor this issue and will revisit it during the 2018 annual review. The DQA has invited the DBA to present its suggestion with supporting data to the DQA. The DQA has also invited other DBAs and Medicaid program administrators to provide input. All of this stakeholder feedback will be incorporated into the next annual review.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures either have only limited baseline scores or will start reporting measures within the next year.

Repeat measurements for two years are available from the CMS CHIPRA Child Health Care Quality Measures reporting. CMS has not released its formal report evaluating trends and changes. However, the data released indicate that in both FFY 2015 and FFY 2016 the median performance was 23.4% in both years across all states reporting the measure. As noted above, 9 additional states reported the measure in FFY 2016 (34 in 2016 versus 25 in 2015). CMS has not reported on improvement among the states who reported the measure in both years.

There also are initial reporting data available from the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after approval by the Dental Quality Alliance and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 461207, 25.41, 25.59, 25.53
2015, 503515, 24.99, 25.18, 24.91

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 76415, 20.17, 22.30, 21.69
2015, 58833, 20.20, 23.14, 22.43

These data also suggest fairly stable rates over the two-year period. However, as noted above, these are initial performance data; additional time may be needed to see improvement within this program. Most measure users are just now getting their quality measurement programs underway.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
No unintended or negative consequences have been identified.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Appendix_Sealants69.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Dental Association on behalf of the Dental Quality Alliance
Co.2 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-
Co.3 **Measure Developer if different from Measure Steward:** American Dental Association on behalf of the Dental Quality Alliance
Co.4 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This project is headed by the DQA through its Measure Development and Maintenance Committee (formerly Research and Development Committee). The following individuals were responsible for executing and overseeing all scientific aspects of this project.

- Craig W. Amundson, DDS, General Dentist, HealthPartners, National Association of Dental Plans. Dr. Amundson serves as chair for the Committee.
- Mark Casey, DDS, MPH, Dental Director, North Carolina Department of Health and Human Services Division of Medical Assistance
- Natalia Chalmers, DDS, PhD, Diplomate, American Board of Pediatric Dentistry, Director, Analytics and Publication, DentaQuest Institute
- Frederick Eichmiller, DDS, Vice President & Science Officer, Delta Dental of Wisconsin
- Chris Farrell, RDH, BSDH, MPA, Oral Health Program Director, Michigan Department of Health and Human Services

This group oversees the maintenance process of the measures. All work of this Committee was distributed for review and formal vote and approval by the entire Dental Quality Alliance. (http://ada.org/dqa) The DQA is made up of representatives from 38 stakeholder organizations.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2013
Ad.3 **Month and Year of most recent revision:** 01, 2017
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2018

Ad.6 **Copyright statement:** 2018 American Dental Association on behalf of the Dental Quality Alliance (DQA) ©. All rights reserved. Use by individuals or other entities for purposes consistent with the DQA's mission and that is not for commercial or other direct revenue generating purposes is permitted without charge.

Ad.7 **Disclaimers:** Dental Quality Alliance measures and related data specifications, developed by the Dental Quality Alliance (DQA), are intended to facilitate quality improvement activities. These Measures are intended to assist stakeholders in enhancing quality of care. These performance Measures are not clinical guidelines and do not establish a standard of care. The DQA has not tested its Measures for all potential applications.

Measures are subject to review and may be revised or rescinded at any time by the DQA. The Measures may not be altered without the prior written approval of the DQA. The DQA shall be acknowledged as the measure steward in any and all references to the measure.

Measures developed by the DQA, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and DQA. Neither the DQA nor its members shall be responsible for any use of these Measures.

THE MEASURES ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND

Limited proprietary coding is contained in the Measure specifications for convenience.

For Proprietary Codes:

The code on Dental Procedures and Nomenclature is published in Current Dental Terminology (CDT), Copyright © 2017 American Dental

Association (ADA). All rights reserved.

This material contains National Uniform Claim Committee (NUCC) Health Care Provider Taxonomy codes (http://www.nucc.org/index.php?option=com_content&view=article&id=14&Itemid=125). Copyright © 2017 American Medical Association. All rights reserved.

Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. The DQA, American Dental Association (ADA), and its members disclaim all liability for use or accuracy of any terminologies or other coding contained in the specifications.

THE SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** In 2008, the Centers for Medicare and Medicaid Services (CMS) asked the ADA to lead the development of a broad coalition of organizations that would lead dentistry to improve the oral health of Americans through quality measurement and quality improvement. The ADA subsequently established the DQA. The DQA is a multi-stakeholder alliance comprised of 38 stakeholders (with organizations as members) from across the oral health community, including federal agencies, third-party payers, professional associations, and an individual member from the general public. The DQA's mission is to advance the field of performance measurement to improve oral health, patient care, and safety through a consensus building process.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2509
Measure Title: Prevention: Sealants for 10-14 Year-Old Children at Elevated Caries Risk, Dental Services
Measure Steward: American Dental Association on behalf of the Dental Quality Alliance
Brief Description of Measure: Percentage of enrolled children in the age category of 10-14 years at "elevated" risk (i.e., "moderate" or "high") who received a sealant on a permanent second molar tooth within the reporting year.
Developer Rationale: Inequalities in oral health status and inadequate use of oral health care services are well documented. Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, L i, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009). Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions.

Evidence-based clinical recommendations recommend that sealants be placed on pits and fissures of children's primary and permanent teeth when it is determined that the tooth, or the patient, is at risk of experiencing caries (Beauchamp et al. 2008). The evidence for sealant effectiveness in permanent molars is stronger than evidence for primary molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013).

The proposed measure, Prevention: Sealants for 10-14 Year-Old Children at Elevated Caries Risk, captures whether children at moderate or high caries risk received a sealant on a permanent second molar tooth. Permanent second molars usually erupt between 10-14 years of age. Thus, this measure addresses both the tooth type on which sealants are placed and the timeliness of care provision. The measure Sealants for 10-14 Year-Old Children allows plans and programs to assess whether children at risk for caries are receiving evidence-based prevention and target performance improvement initiatives accordingly.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

Numerator Statement: Unduplicated number of enrolled children age 10-14 years at "elevated" risk (i.e., "moderate" or "high") who received a sealant on a permanent second molar tooth as a dental service.
Denominator Statement: Unduplicated number of enrolled children age 10-14 years who are at "elevated" risk (i.e., "moderate" or "high")
Denominator Exclusions: Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Sep 18, 2014 **Most Recent Endorsement Date:** Sep 18, 2014

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (sealants) as well as the specific tooth type for which the evidence is the strongest (permanent molar) and the timing of sealant placement to maximize effectiveness (shortly after eruption – 10-14 years of age for permanent second molars). "Caries Prevention: Sealants should be placed on pits and fissures of children's and adolescents' permanent teeth when it is determined that the tooth, or the patient, is at risk of developing caries." (Beauchamp et al. 2008, p. 263, Table 3)
- Grade/Strength of Recommendation: B which is defined as: "Directly based on category II evidence or extrapolated recommendation for category I evidence." (Beauchamp 2008, pp. 261, 263, Tables 1,2, 3)

Citation:

Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268. Available at: http://jada.ada.org/content/139/3/257.full.

Changes to evidence from last review

- ☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
- ☒ **The developer provided updated evidence for this measure:**

Updates:

- A recent Cochrane Review on the effectiveness of sealants brings together all the evidence on this topic. The conclusions of this new review continue to support the recommendations of the ADA Sealant Guideline (Note: the ADA is currently updating this guideline).

Citation:

Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

Question for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review, but does note a recent Cochrane review collated all evidence and reached the same conclusions that supported the original guideline. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm
Process measure based on systematic review (Box 3)→ QQC presented (Box 4)→ Contains Quantity: High (7 systematic reviews and 14 individual clinical studies) Quality: High, Consistency: Moderate → Rate as High

Preliminary rating for evidence: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**
RATIONALE: The quality of the evidence is high (systematic reviews of randomized controlled trials), for sealants placed on permanent molars of children and adolescents. The evidence directly pertains to both the measure focus and the measure target population.

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer used data from five sources and refers to "program" level information and "plan" level information (Texas Medicaid, Texas CHIP, Florida CHIP, and Florida Medicaid programs as well as national commercial data from Dental Service of Massachusetts, Inc.). The developer presented the total number of children enrolled in each program/plan. In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011.
- The data source and sample size are sufficient to assess gaps in performance. The performance range of **8% to 13% in CY 2010** (year in which data were available for all five programs) indicate low sealant placement prevalence rates as well as variations in sealant prevalence across programs. Data from the Centers for Medicare and Medicaid Services (CMS) indicate significant variation among state Medicaid programs, ranging from 6% to 22% of children 10-14 years old, who received a sealant on a permanent molar tooth (CMS-416 data, FY 2011).
- The developer did not provide more recent performance data, stating that due to the start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting.

Disparities

- The developer found disparities based by age, geographic location, and race/ethnicity. In addition, it also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), CHIP dental plan, Medicaid program type, commercial product line, and preferred language for program communications. It detected disparities based on each of these various factors, but data on all of these characteristics were not consistently available for all programs so it presented disparities data on those characteristics that were most consistently available and had the greatest standardization (i.e. race/ethnicity and geographic location).

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**
RATIONALE: There appears to be significant variation in performance across plans/programs.

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Staff Scientific Acceptability Logic

*The original testing was submit as permitted by NQF.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure relies on **standard data elements in administrative claims data** (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes.

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**
RATIONALE: All data elements are in defined fields in electronic claims.

Committee pre-evaluation comments
Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- Texas Health and Human Services Commission: Texas Medicaid/CHIP Pay-for-Quality (P4Q) program. https://hhs.texas.gov/sites/default/files//documents/lawsregulations/ handbooks/umcm/6-2-15.pdf

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- In 2016, the Dental Quality Alliance (DQA) expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from six state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from eight dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

Additional Feedback:

- A dental benefits administrator (DBA) suggested that the DQA consider adding patient exclusions to the measure. The DQA considered exclusions previously during initial measure development and during annual reviews. Exclusions were not included due to concerns about the introduction of biased measurement, increased measurement complexity, and adversely affecting implementation feasibility. However, the DQA continues to monitor this issue and will revisit it during the 2018 annual review. The DQA has invited the DBA to present its suggestion with supporting data to the DQA. The DQA also has invited other DBAs and Medicaid program administrators to provide input. All of this stakeholder feedback will be incorporated into the next annual review.

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- The developer did not provide more recent performance data, stating that due to the start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting.

The developer provides data from the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after it was approved by the DQA and before NQF endorsement, as follows:

Texas Medicaid
2014, 475976, 16.78, 17.10, 16.59
2015, 527493, 16.63, 16.48, 16.90

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 102148, 12.59, 14.08, 12.96
2015, 70216, 12.59, 13.90, 14.28

The developer notes that these data suggest fairly stable rates over the two-year period—i.e., improvement is not noted. However, as noted above, these are initial performance data for one program, and additional time is likely to be needed to see improvement because most measure users are just now getting their quality measurement programs underway.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

No unintended or negative consequences were identified by the developer.
Preliminary rating for Usability and use: ☐ High ☒ Moderate ☐ Low ☐ Insufficient
Committee pre-evaluation comments **Criteria 4: Usability and Use**

Criterion 5: Related and Competing Measures
Related or competing measures • N/A **Harmonization** • N/A
Committee pre-evaluation comments **Criterion 5: Related and Competing Measures**

Public and member comments

Staff Scientific Acceptability Logic

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

☒Yes (go to Question #4)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☒Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

   ☐Yes (go to Question #5)

   ☒No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☐Yes (go to Question #6)

   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?

   ☐High (go to Question #8)

   ☐Moderate (go to Question #8)

   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?

   ☐Yes (go to Question #8)

   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

   ☒Yes (go to Question #9)

   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☒Yes (go to Question #10)

   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and <u>all</u> testing results:

☐High (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?
   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.
   ☒Yes (go to Question #2)
   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?
   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?
   ☐Yes (please explain below then go to Question #3)
   ☐No (go to Question #3)
   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?
   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*
   ☐Yes (please explain below then go to Question #4)
   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
   ☐Yes (please explain below then go to Question #5)
   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
   ☐Yes (please explain below then go to Question #6)
   ☒No (go to Question #6)

☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☐Yes (go to Question #11)
    ☒No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☐Yes (go to Question #12)

☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
    ☐High (go to Question #14)
    ☐Moderate (go to Question #14)
    ☐Low (please explain below then go to Question #13)
    ☐Insufficient

13. Was other validity testing reported?
    ☒Yes (go to Question #14)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?
    TIPS: Prior validity studies of the same data elements may be submitted
    ☒Yes (go to Question #15)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*
    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
    ☒Yes (go to Question #16)
    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
    ☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
    ☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Prevention: Dental Sealants for 10-14 Year-Old Children at Elevated Caries Risk

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 2/10/2014

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.

- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☐ Health outcome:

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome:

X☐ Process: Receipt of evidence-based preventive dental services – sealants on permanent molars - during the reporting period

☐ Structure:

☐ Other:

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not applicable.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

Not applicable.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Sealants for 10-14 Year-Old Children at Elevated Caries Risk indicates the percentage of children at moderate to high risk for caries who received a sealant on a second permanent molar. Evidence-based clinical recommendations recommend that sealants be placed on pits and fissures of children's primary and permanent teeth when it is determined that the tooth, or the patient, is at risk of experiencing caries, with greater evidence of effectiveness in permanent molars compared to primary molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013). This measure directly reflects evidence based guidelines regarding an effective caries prevention measure (sealants) as well as the specific tooth type for which the evidence is the strongest (permanent molar) and the timing of sealant placement to maximize effectiveness (shortly after eruption – 10-14 years of age for permanent second molars). As described in 1b1 (Importance), dental caries is the most common chronic disease in children in the U.S. and a significant percentage of children have untreated dental caries. Dental decay causes significant short- and long-term adverse consequences for children's health and functioning. As detailed below, timely placement of sealants on permanent second molars have demonstrated effectiveness in reducing caries among children, thereby improving oral health, overall health, and overall well-being.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268. Available at: http://jada.ada.org/content/139/3/257.full.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

"Caries Prevention: Sealants should be placed on pits and fissures of **children's** and **adolescents'** permanent teeth when it is determined that the tooth, or the patient, is at risk of developing caries." (Beauchamp et al. 2008, p. 263, Table 3)

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade/Strength of Recommendation: B which is defined as: "Directly based on category II evidence or extrapolated recommendation for category I evidence." (Beauchamp 2008, pp. 261, 263, Tables 1,2, 3)
[See grades for strength of evidence in section 1a7.]

Grading system adapted from: Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. BMJ 1999;318(7183):593-596.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

A: Directly based on category I evidence

B: Directly based on category II evidence or extrapolated recommendation from category I evidence

C: Directly based on category III evidence or extrapolated recommendation from category I or II evidence

D: Directly based on category IV evidence or extrapolated recommendation from category I, II or III evidence

Grading system adapted from: Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. BMJ 1999;318(7183):593-596.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same as that provided for the guidelines provided in 1a.4.1.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ **X**Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable.

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

Not applicable.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable.

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Not applicable.

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable.

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268. Available at: http://jada.ada.org/content/139/3/257.full.

Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Not applicable.

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The following four clinical questions were addressed:

- "Under what circumstances should sealants be placed to prevent caries?"
- "Does placing sealants over early (noncavitated) lesions prevent progression of the lesions?"
- "Are there conditions that favor the placement of resin-based versus glass ionomer cement sealants in terms of retention or caries prevention?"
- "Are there any techniques that could improve sealants' retention and effectiveness in caries prevention?" (Beauchamp et al. 2008, pp. 259-260)

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

"Caries Prevention: Sealants should be placed on pits and fissures of **children's** and **adolescents'** permanent teeth when it is determined that the tooth, or the patient, is at risk of developing caries." (Beauchamp et al. 2008, p. 263, Table 3)

Grade: The evidence grade is **IA** which is defined as: "Evidence from systematic reviews of randomized controlled trials" (Beauchamp 2008, pp. 261, 263, Tables 1, 3). Grading system adapted from: Shekelle et al. (1999) cited in 1a.4.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Ia: Evidence from systematic reviews of randomized controlled trials

Ib: Evidence from at least one randomized controlled trial

IIa: Evidence from at least one controlled study without randomization

IIb: Evidence from at least one other type of quasiexperimental study, such as time series analysis or studies in which the unit of analysis is not the individual

III: Evidence from nonexperimental descriptive studies, such as comparative studies, correlation studies, cohort studies and case-control studies

IV: Evidence from expert committee reports or opinions or clinical experience of respected authorities

(Beauchamp et al. 2008, p. 261) Grading system adapted from: Shekelle et al. (1999).

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Literature studies for sealants were conducted to identify all systematic reviews through Oct. 4, 2006. To ensure new clinical studies published since the search within each review were included within the guideline development effort, additional searches were conducted for clinical trials until September 2006.

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

7 systematic reviews and 14 individual clinical studies were reviewed with respect to the clinical questions identified. The evidence guidelines do not provide summary data regarding the number of studies by type of study. (Beauchamp 2008, p. 260)

However, the guidelines provide the following details regarding the strength and quality of the evidence related to sealants for caries prevention:

Evidence Grade Ia (systematic reviews of randomized controlled trials)

Supports the following evidence statements based on the evidence review by the expert panel:

- "Placement of resin-based sealants on the permanent molars of children and adolescents is effective for caries reduction." (Beauchamp 2008, p. 260)
- "Reduction of caries incidence in children and adolescents after placement of resin-based sealants ranges from 86 percent at one year to 78.6 percent at two years and 58.6 percent at four years." (Beauchamp 2008, p. 260)

Studies with evidence grade of Ia cited:

Ahovuo-Saloranta A, Hiiri A, Nordblad A, Worthington H, Mäkelä M. Pit and fissure sealants for preventing dental decay in the permanent teeth of children and adolescents. Cochrane Database Syst Rev 2004(3):CD001830.

Llodra JC, Bravo M, Delgado-Rodriguez M, Baca P, Galvez R. Factors influencing the effectiveness of sealants: a meta-analysis. Community Dent Oral Epidemiol 1993;21(5):261-268.

Evidence Grade Ib (evidence from at least one randomized controlled trial)

Supports the following evidence statements based on the evidence review by the expert panel:

- "Sealants are effective in reducing occlusal caries incidence in permanent first molars of children, with caries reductions of 76.3 percent at four years, when sealants were reapplied as needed. Caries reduction was 65 percent at nine years from initial treatment, with no reapplication during the last five years." (Beauchamp 2008, p. 261)

Studies with evidence grade of Ib cited:

Bravo M, Montero J, Bravo JJ, Baca P, Llodra JC. Sealant and fluoride varnish in caries: a randomized trial. J Dent Res 2005;84(12):1138-1143.

Evidence Grade III (evidence from nonexperimental descriptive studies, such as comparative studies, correlation studies, cohort studies and case control studies)

Supports the following evidence statements based on the evidence review by the expert panel:

- "There is consistent evidence from private dental insurance and Medicaid databases that placement of sealants on first and second permanent molars in children and adolescents is associated with reductions in the subsequent provision of restorative service." (Beauchamp 2008, p. 261)

- "Evidence from Medicaid claims data for children who were continuously enrolled for four years indicates that sealed permanent molars are less likely to receive restorative treatment, that the time between receiving sealants and receiving restorative treatment is greater, and that the restorations were less extensive than those in permanent molars that were unsealed." (Beauchamp 2008, p. 261)

Studies with evidence grade of III cited:

Bhuridej P, Damiano PC, Kuthy RA, et al. Natural history of treatment outcomes of permanent first molars: a study of sealant effectiveness. JADA 2005;136(9):1265-1272.

Dennison JB, Straffon LH, Smith RC. Effectiveness of sealant treatment over five years in an insured population. JADA 2000;131(5):597-605.

Hotuman E, Rølling I, Poulsen S. Fissure sealants in a group of 3-4-year-old children. Int J Paediatr Dent 1998;8(2):159-160.

Weintraub JA, Stearns SC, Rozier RG, Huang CC. Treatment outcomes and costs of dental sealants among children enrolled in Medicaid. Am J Public Health 2001;91(11):1877-1881.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The quality of the evidence is high, grades of Ia (systematic reviews of randomized controlled trials), for sealants placed on permanent molars of children and adolescents.

The evidence directly pertains to both the measure focus and the measure target population.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Meta-analyses were not conducted as part of the evidence review. Please see the response in 1a.7.5. regarding the identified benefits and associated strength of evidence. However, a more recent Cochrane Review published in 2013 by Ahovuo-Saloranta et al. brings together all the evidence in a quantitative manner. More information from this review is provided below in Section 1.a.7.9

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Harms were not evaluated as part of this systematic review. However this question was addressed in a recent Cochrane Review on the effectiveness of sealants (Ahovuo-Saloranta et al. 2013), and it was noted: "Only two studies (Bravo 2005; Liu 2012) assessed side effects of the sealants. No adverse effects were detected or reported by patients included in the studies."

Citations:

Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

Bravo M, Montero J, Bravo JJ, Baca P, Llodra JC. Sealant and fluoride varnish in caries: a randomized trial. Journal of Dental Research 2005;84(12):1138-43.

Liu BY, Lo ECM, Chu CH, Lin HC. Randomized trial on fluorides and sealants for fissure caries prevention. Journal of Dental Research 2012;91(8):753-8.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

A recent Cochrane Review on the effectiveness of sealants brings together all the evidence on this topic. The conclusions of this new review continue to support the recommendations of the ADA Sealant Guideline (Note: the ADA is currently updating this guideline). The summary of findings from the Cochrane review appears below

SUMMARY OF FINDINGS FOR THE MAIN COMPARISON *[Explanation]*

Resin-based sealant compared to control without sealant for preventing dental caries

Patient or population: Children and adolescents
Settings: Sealant applications for school children in USA, Canada, China & Colombia
Intervention: Resin-based sealant applications on occlusal tooth surfaces of permanent molars
Comparison: No sealant application

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	Number of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk	Corresponding risk				
	Control teeth	Sealed teeth				
Dentine caries in permanent molars Follow-up: 2 years	Incidence of carious first molars (40%) 400 per 1000[1]	Incidence of carious first molars (6.3%) 63 per 1000 (38 to 96)	OR 0.12 (0.07 to 0.19) [2]	1259 children randomised & 1066 evaluated after 2 years (6 studies [3,4,5])	⊕⊕⊕○ **moderate**	Benefits of resin-sealant maintained up to at least 48 months of follow-up[6]
	Incidence of carious first molars (70%) 700 per 1000[1]	Incidence of carious first molars (19%) 190 per 1000 (122 to 272)	OR 0.12 (0.07 to 0.19) [2]	1259 children randomised & 1066 evaluated after 2 years (6 studies [3,4,5])	⊕⊕⊕○ **moderate**	Benefits of resin-based sealant maintained up to at least 48 months of follow-up[6]

CI: confidence interval; **OR:** odds ratio

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] The incidence of carious control teeth in the five split-mouth trials included in this comparison ranged from 37% to 69% (studies published between 1976 and 1979). We have shown the effect of sealants at each end of this range. These studies did not give information on the baseline caries prevalence of the children.
The sixth study included in this meta-analysis (parallel group study published in 2012) reported clearly lower incidence of carious first molars than the five split-mouth studies. In sealant group, carious first molars were detected in 9 out of 121 children (7.4%) (11 carious teeth out of 367 sealed teeth) and in placebo group in 21 out of 124 children (17%) (28 carious teeth out of 379 placebo teeth). Caries prevalence: mean baseline dmft level of 3.4.
[2] There was considerable heterogeneity in this estimate (I^2 = 77% P = 0.0007) but all of the trials showed a statistically significant effect favouring sealants.
[3] Six studies at low risk of bias for the four key domains of allocation concealment, incomplete outcome data, selective reporting and baseline comparability of the groups.
[4]All studies recruited children aged 5-10 years. Three studies conducted in areas with fluoridated water, two studies stated water was not fluoridated and the remaining one study did not report whether water supplies were fluoridated.
[5] Five trials were published between 1976 and 1979 and one in 2012. One further parallel group trial from Thailand at unclear risk of bias reporting DFS increment published in 1995 also found a benefit in favour of resin-based sealant (mean difference in DFS increment -0.65, 95% CI -0.83 to -0.47, 276 children evaluated).
[6] The benefit associated with sealant use is maintained at all of the follow-up estimates (up to 9 years) though the number of studies and the number of children available for evaluation reduced markedly over this period (e.g. at 48 to 54 months of follow-up odds ratio 0.21, 95% CI 0.16 to 0.28, two studies at low risk of bias and two studies at high risk of bias, 482 children evaluated; risk ratio 0.24, 95% CI 0.12 to 0.45, one study at unclear risk of bias, 203 children evaluated).

Citations

Ahovuo-Saloranta A, Forss H, Walsh T, Hiiri A, Nordblad A, Mäkelä M, Worthington HV. Sealants for preventing dental decay in the permanent teeth. Cochrane Database Syst Rev. 2013 Mar 28;3:CD001830. doi: 10.1002/14651858.CD001830.pub4.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

Not applicable.

1a.8.2. Provide the citation and summary for each piece of evidence.

Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***
1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form 4_Evidence_10-14.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?
Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.
No

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*
If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
Inequalities in oral health status and inadequate use of oral health care services are well documented. Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, L i, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009). Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions.

Evidence-based clinical recommendations recommend that sealants be placed on pits and fissures of children's primary and permanent teeth when it is determined that the tooth, or the patient, is at risk of experiencing caries (Beauchamp et al. 2008). The evidence for sealant effectiveness in permanent molars is stronger than evidence for primary molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013).

The proposed measure, Prevention: Sealants for 10-14 Year-Old Children at Elevated Caries Risk, captures whether children at moderate or high caries risk received a sealant on a permanent second molar tooth. Permanent second molars usually erupt between 10-14 years of age. Thus, this measure addresses both the tooth type on which sealants are placed and the timeliness of care provision. The measure Sealants for 10-14 Year-Old Children allows plans and programs to assess whether children at risk for caries are receiving evidence-based prevention and target performance improvement initiatives accordingly.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Below are the testing data and results that met scientific acceptability criteria for endorsement. Because there were no changes in the data source, level of analysis or setting, additional testing has not been conducted.

Data Sources:

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. Data from calendar years 2010 and 2011 were used for all programs except Florida Medicaid. Full-year data for CY 2011 were not available for Florida Medicaid. Therefore, we report only CY 2010 data for Florida Medicaid.

In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011. [Technically, there were three plans represented in the data because Texas CHIP was served by a single dental plan. Since the program=plan in that case, we included it in the "program" level data.]

Below we provide summary data for each of the five programs and two plans individually.

Programs

Our source data for the testing prior to applying the denominator age criteria of 10-14 years old included children 0-20 years in each program. The number of children ages 0-20 years enrolled at least one month in each program were as follows:

Texas Medicaid, 2011: 3,544,247
Texas Medicaid, 2010: 3,393,963
Texas CHIP, 2011: 842,454
Texas CHIP, 2010: 786,070
Florida CHIP, 2011: 317,146
Florida CHIP, 2010: 315,975
Commercial, 2011: 184,152
Commercial, 2010: 189,968
Florida Medicaid, 2010: 2,068,670

Within these programs, we had claims data available in both years for two dental managed care plans in Florida CHIP. We also report rates for those two plans separately.

Plan 1, 2010: 77,255
Plan 2, 2010: 116,388
Plan 1, 2011: 140,986
Plan 2, 2011: 168,191

The number of children in the age range of 10-14 years specifically was:

Texas Medicaid, 2011: 732,230
Texas Medicaid, 2010: 678,393
Texas CHIP, 2011: 283,104
Texas CHIP, 2010: 263,541
Florida CHIP, 2011: 125,095
Florida CHIP, 2010: 124,914
Commercial, 2011: 49,789
Commercial, 2010: 51,634
Florida Medicaid, 2010: 426,206
Plan 1, 2010: 29,214
Plan 2, 2010: 45,652
Plan 1, 2011: 55,456
Plan 2, 2011: 66,535

Data 1b.2. Performance Scores for Dental Sealants for 10-14 Year-Olds at Elevated Risk

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

Program 1, CY 2011:	11.08%	(	0.1108	,	0.0005	,	0.1099	,	0.1117)
Program 2, CY 2011:	10.59%	(	0.1059	,	0.0010	,	0.1039	,	0.1079)
Program 3, CY 2011:	10.58%	(	0.1058	,	0.0018	,	0.1022	,	0.1094)
Program 4, CY 2011:	10.84%	(	0.1084	,	0.0026	,	0.1034	,	0.1134)
Program 1, CY 2010:	10.48%	(	0.1048	,	0.0005	,	0.1038	,	0.1058)
Program 2, CY 2010:	7.67%	(	0.0767	,	0.0009	,	0.0749	,	0.0785)
Program 3, CY 2010:	10.36%	(	0.1036	,	0.0018	,	0.1000	,	0.1072)
Program 4, CY 2010:	12.70%	(	0.1270	,	0.0027	,	0.1217	,	0.1323)
Program 5, CY 2010:	8.44%	(	0.0844	,	0.0011	,	0.0823	,	0.0865)

Plan 1, CY 2011:	9.64%	(	0.0964	,	0.0027	,	0.0911	,	0.1017)
Plan 2, CY 2011:	11.06%	(	0.1106	,	0.0025	,	0.1056	,	0.1156)
Plan 1, CY 2010:	10.10%	(	0.1010	,	0.0045	,	0.0923	,	0.1097)
Plan 2, CY 2010 :	10.05%	(	0.1005	,	0.0032	,	0.0943	,	0.1067)

The measure rate range of 8% to 13% in CY 2010 (year in which data were available for all five programs) indicate low sealant placement prevalence rates as well as variations in sealant prevalence across programs.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

The measure testing findings are consistent with other data indicating that there are significant variations in the percentage of children who received sealants. Data from the Centers for Medicare and Medicaid Services indicate significant variation among state Medicaid programs, ranging from 6% to 22% of children 10-14 years old, who received a sealant on a permanent molar tooth (CMS-416 data, FY 2011).

[Complete citations provided in 1c4 and in Evidence Submission Form Template.]

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The same data sources were used as described in 1b.2. The data below summarizes performance data by geographic location and race/ethnicity for CY 2011 (CY 2010 for one program) with the p-values from chi-square tests used to detect whether there were statistically significant differences in performance between groups. Disparities by geographic location were detected for three programs. Statistically significant difference in performance by race and ethnicity also were detected in the two programs for which there were race/ethnicity data. In addition, we also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), Medicaid program type, CHIP dental plan, commercial product line, and preferred language for program communications. We additionally detected disparities by health status, dental plan and Medicaid program type, but data on all of these characteristics were not consistently available for all programs so we are presenting disparities data on those characteristics that were most consistently available and had the greatest standardization

Data1b.4. Disparities in Performance by Geographic Location and Race/Ethnicity

PROGRAM 1	
Overall performance score:	11.08%
Scores by Geographic Location	
Urban:	11.36%
Rural:	9.32%
p-value from Chi-square test:	<.0001
Scores by Race	
Non-Hispanic White:	8.54%
Non-Hispanic Black:	10.95%

Hispanic: 11.85%
p-value from Chi-square test <.0001

PROGRAM 2
Overall performance score: 10.59%
Scores by Geographic Location
Urban: 10.73%
Rural: 9.77%
p-value from Chi-square test: 0.0029
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 3
Overall performance score: 10.58%
Scores by Geographic Location
Urban: 10.44%
Rural: 12.99%
p-value from Chi-square test: 0.0008
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 4
Overall performance score: 10.84%
Scores by Geographic Location
Urban: 10.86%
Rural: 10.33%
p-value from Chi-square test: 0.7002
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 5
Overall performance score: 8.44%
Scores by Geographic Location
Urban: 8.40%
Rural: 9.09%
p-value from Chi-square test: 0.1546
Scores by Race
Non-Hispanic White: 8.16%
Non-Hispanic Black: 8.93%
Hispanic: 8.21%
p-value from Chi-square test 0.0054

Note: N/A for race/ethnicity indicates that those programs did not collect race/ethnicity data or had high rates of missing data .

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

With respect to preventive dental services in general, there are documented disparities. Using data from the National Survey of Children's Health, Edelstein and Chinn (2009) noted disparities in access to preventive dental services by race and income: "Stepwise disparities in access to preventive dental services are evident by race and income in ways that parallel Medical Expenditure Panel Survey findings. White parents report higher use of preventive dental services than do black or Hispanic parents (77%, 66%, and 61%, respectively). Poor parents report less use of services than do low income, middle class, and higher-income parents (58%, 66%, 77%, and 82%, respectively)" (Edelstein & Chinn, 2009, p.418). A recent analysis by Bouchery (2013) of the Medicaid Analytic eXtract files for nine states found variations in the percentage of children receiving a preventive dental visit by age, race and ethnicity, and geographic area. Specifically, relative to the reference group of 9 year olds, the percentage point change in the probability of having a dental preventive services was -27.6 for 3 years old; -8.6 for 6 years, -2.2 for 12 years and -15.4 for 15 years (all significant at $p<0.0001$); relative to the reference group of white, non-Hispanic, the percentage point change was -1.8 for black non-Hispanic and 7.8 for Hispanic ($p<0.0001$ for both); relative to the reference group of small metro area, the percentage point change was 5.9 for large metro area ($p<0.0001$).

In addition, there are documented disparities in dental sealant receipt specifically. For example, using data from the National Health and Nutrition Examination Survey, researchers at the National Center for Health Statistics identified variations in dental sealant prevalence among children by age, race, ethnicity, and poverty level (Dye, Li, and Thorton-Evans 2012). Specifically: "Dental sealant prevalence was lower among children [6-9 years] living at or below 100% of the federal poverty level (26%) compared with children living above the poverty level (34%). A similar pattern was found among adolescents aged 13–15, but the difference was not statistically significant. Dental sealant prevalence was significantly lower for non-Hispanic black adolescents (32%) compared with non-Hispanic white adolescents (56%), among those aged 13–15" (Dye, Li, and Thorton-Evans 2012, p. 2).

Sources

Bouchery, E. 2013. "Utilization of Dental Services among Medicaid-Enrolled Children." Medicare & Medicaid Research Review. 3(3) E1-16. Available at: https://www.cms.gov/mmrr/Downloads/MMRR2013_003_03_b04.pdf.

Dietrich, T., C. Culler, R. Garcia, and M. M. Henshaw. 2008. Racial and ethnic disparities in children's oral health: The National Survey of Children's Health. Journal of the American Dental Association 139(11):1507-1517.

Dye BA, Li X, Thorton-Evans G. Oral health disparities as determined by selected healthy people 2020 oral health objectives for the United States, 2009-2010. NCHS Data Brief 2012(104):1-8.U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research.

Edelstein, B. L. and C. H. Chinn. 2009. "Update on Disparities in Oral Health and Access to Dental Care for America´s Children." Acad Pediatr 9(6): 415-9.

Institute of Medicine (U.S.). Committee on an Oral Health Initiative. Advancing oral health in America. Washington, D.C.: National Academies Press; 2011.

Institute of Medicine and National Research Council. Improving access to oral health care for vulnerable and underserved populations. Washington, D.C.: National Academies Press; 2011.

Kenney, G. M., J. R. McFeeters, and J. Y. Yee. 2005. Preventive dental care and unmet dental needs among low-income children. American Journal of Public Health 95(8):1360-1366.

Lewis, C., W. Mouradian, R. Slayton, and A. Williams. 2007. Dental insurance and its impact on preventative dental care visits for U.S. children. Journal of the American Dental Association 138(3):369-380.

Oral Health in America: a report of the Surgeon General. Rockville, Md.: U.S. Public Health Service, Dept. of Health and Human Services; 2000.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Dental

De.6. **Non-Condition Specific***(check all the areas that apply):*
Access to Care, Disparities Sensitive, Health and Functional Status : Change, Health and Functional Status : Total Health, Primary Prevention

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, Populations at Risk

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.ada.org/~/media/ADA/Science%20and%20Research/Files/DQA_2018_Dental_Services_Sealants_10-14_years.pdf?la=en

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
1. No changes to the measure specifications
2. Measure specification website updated to be more user friendly

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Unduplicated number of enrolled children age 10-14 years at “elevated” risk (i.e., “moderate” or “high”) who received a sealant on a permanent second molar tooth as a dental service.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in*

required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see Section S14

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Unduplicated number of enrolled children age 10-14 years who are at "elevated" risk (i.e., "moderate" or "high")

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see Section S14.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
There are no other exclusions than those described above.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
There are no stratifications for this measure.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Sealants for 10-14 year olds - Calculation for Children at Elevated Caries Risk

1. Use administrative enrollment and claims data for a single year. When using claims data to determine service receipt, include both paid and unpaid claims (including pending, suspended, and denied claims).

2. Check if the enrollee meets age criteria at the last day of the reporting year:
a. If child is >= 10 and <= 14, then proceed to next step.
b. If age criterion is not met or there are missing or invalid field codes (e.g., date of birth), then STOP processing. This enrollee does not get counted.

3. Check if subject is continuously enrolled for at least 180 days during the reporting year:

a. If subject meets continuous enrollment criterion, then proceed to next step.
b. If subject does not meet enrollment criterion, then STOP processing. This enrollee does not get counted.

YOU NOW HAVE THE COUNT OF THOSE WHO MEET THE AGE AND ENROLLMENT CRITERIA

4. Check if subject is at "elevated risk":
a. If subject meets ANY of the following criteria, then include in denominator:
i. the subject has a CDT Code among those in Table 1 in the reporting year,
OR
ii. the subject has a CDT Code among those in Table 1 in any of the three years prior to the reporting year, (NOTE: The subject does not need to be enrolled in any of the prior three years for the denominator enrollment criteria; this is a "look back" for enrollees who do have claims experience in any of the prior three years.)
OR
iii. the subject has a visit with a CDT code = (D0602 or D0603) in the reporting year.
b. If the subject does not meet any of the above criteria for elevated risk, then STOP processing. This enrollee will not be included in the measure denominator.

YOU NOW HAVE THE DENOMINATOR (DEN): Enrollees who are at "elevated risk"

5. Check if subject received a sealant as a dental service during the reporting year:
a. If [CDT CODE] = D1351, and;
b. If [RENDERING PROVIDER TAXONOMY] code = any of the NUCC maintained Provider Taxonomy Codes in Table 2 below, then proceed to next step.
c. If both a AND b are not met, then the service was not a "dental service"; STOP processing. This enrollee is already included in the denominator but will not be included in the numerator.

Note: In this step, all claims with missing or invalid CDT CODE, missing or invalid NUCC maintained Provider Taxonomy Codes, or NUCC maintained Provider Taxonomy Codes that do not appear in Table 2 should not be included in the numerator.

6. Check if sealant was placed on a permanent second molar:
a. If [TOOTH-NUMBER] = 2, 15, 18, 31 then include in numerator; STOP processing.
b. If not, then service was not provided for the second permanent molar; STOP processing. This enrollee is already included in the denominator but will not be included in the numerator.

YOU NOW HAVE NUMERATOR (NUM) COUNT: Enrollees at "elevated risk" who received sealants on a permanent second molar as a dental service

7. Report
a. Unduplicated number of enrollees in numerator
b. Unduplicated number of enrollees in each denominator
c. Measure rate (NUM/DEN)

Table 1: CDT Codes to identify "elevated risk"

D2140	D2394	D2630	D2720	D2791	D3120
D2150	D2410	D2642	D2721	D2792	D3220
D2160	D2420	D2643	D2722	D2794	D3221
D2161	D2430	D2644	D2740	D2799	D3222
D2330	D2510	D2650	D2750	D2930	D3230
D2331	D2520	D2651	D2751	D2931	D3240
D2332	D2530	D2652	D2752	D2932	D3310
D2335	D2542	D2662	D2780	D2933	D3320
D2390	D2543	D2663	D2781	D2934	D3330
D2391	D2544	D2664	D2782	D2940	D2941
D2392	D2610	D2710	D2783	D2950	D1354
D2393	D2620	D2712	D2790	D3110	

Table 2: NUCC maintained Provider Taxonomy Codes classified as "Dental Service"*

122300000X	1223P0106X	1223X0008X	261QF0400X
1223D0001X	1223P0221X	1223X0400X	261QR1300X
1223D0004X	1223P0300X	124Q00000X+	125Q00000X
1223E0200X	1223P0700X	125J00000X	
1223G0001X	1223S0112X	125K00000X	

*Services provided by County Health Department dental clinics may also be included as "dental" services.
+Only dental hygienists who provide services under the supervision of a dentist should be classified as "dental" services. Services provided by independently practicing dental hygienists should be classified as "oral health" services and are not applicable for this measure.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
Not applicable.

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
Not applicable.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
Not applicable.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable

2. Validity – See attached Measure Testing Submission Form
5_Testing_10-1.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Prevention: Dental Sealants for 10-14 Year-Old Children at Elevated Caries Risk

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.

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] 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] 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]

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]

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; [14,15] 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] 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
☒ 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:	☐ other:

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
   A. 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
   A. 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
   A. 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
   A. 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
   A. Size: # Members 0-20 years, CY 2010: 2,068,670;
   B. 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

2) FL CHIP – Plan 2
   A. 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 <u>patients</u> 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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2011					
	Program 1	Program 2	Program 3	Program 4	Plan 1	Plan 2
Total Number Patients	3,544,247	842,454	317,146	184,152	140,986	168,191
Age Group Distribution						
Age <1 years	7.05%	0.11%	N/A	1.54%	N/A	N/A
Age 1-2 years	14.32%	5.34%	N/A	5.75%	N/A	N/A
Age 3-5 years	19.46%	11.70%	3.81%	12.68%	4.12%	3.60%
Age 6-7 years	11.21%	12.30%	13.05%	9.57%	13.71%	12.55%
Age 8-9 years	9.85%	14.40%	15.00%	10.18%	15.76%	14.41%
Age 10-11 years	9.03%	14.03%	15.71%	10.55%	16.27%	15.25%
Age 12-14 years	11.63%	19.57%	23.73%	16.09%	23.06%	24.31%
Age 15-18 years	13.19%	22.54%	28.70%	22.13%	27.08%	29.88%
Age 19-20 years	4.27%	N/A	N/A	11.50%	N/A	N/A
Geographic Location						
Urban	83.63%	84.33%	92.94%	95.95%	93.01%	92.91%
Rural	15.15%	14.61%	5.02%	3.86%	4.83%	5.15%
Missing	1.22%	1.06%	2.04%	0.19%	2.16%	1.94%
Race and Ethnicity						
Non-Hispanic White	17.36%	N/A	N/A	N/A	N/A	N/A
Non-Hispanic Black	15.08%	N/A	N/A	N/A	N/A	N/A
Hispanic	58.07%	N/A	N/A	N/A	N/A	N/A
Other & Unknown	9.49%	N/A	N/A	N/A	N/A	N/A

Table 1.6B, Patient Characteristics, 10-14 Years Old (Age Range Targeted by Measure), 2011

	Descriptive Characteristics of Individuals 10-14 Years Enrolled at Least One Month, CY 2011					
	Program 1	Program 2	Program 3	Program 4	Plan 1	Plan 2
Total Number Patients	732,230	283,104	125,095	49,789	55,456	66,535
Age Group Distribution						
Age 10-11 years	43.69%	41.76%	39.84%	39.60%	41.38%	38.55%
Age 12-14 years	56.31%	58.24%	60.16%	60.40%	58.62%	61.45%
Geographic Location						
Urban	83.92%	84.31%	93.14%	95.86%	93.24%	93.07%
Rural	15.23%	14.59%	5.05%	3.97%	4.82%	5.20%
Missing	0.84%	1.10%	1.81%	0.17%	1.94%	1.72%
Race and Ethnicity						
Non-Hispanic White	17.09%	N/A	N/A	N/A	N/A	N/A
Non-Hispanic Black	16.35%	N/A	N/A	N/A	N/A	N/A
Hispanic	59.18%	N/A	N/A	N/A	N/A	N/A
Other & Unknown	7.37%	N/A	N/A	N/A	N/A	N/A

Table 1.6C, Patient Characteristics, 0-20 Years Old, 2010

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2010						
	Program 1	Program 2	Program 3	Program 4	Program 5	Plan 1	Plan 2
Total Number Patients	3,393,963	786,070	315,975	189,968	2,068,670	77,255	116,388
Age Group Distribution							
Age <1 years	7.35%	0.15%	N/A	1.45%	6.05%	N/A	N/A
Age 1-2 years	15.16%	5.37%	N/A	5.67%	14.23%	N/A	N/A
Age 3-5 years	19.48%	11.69%	3.64%	12.73%	19.26%	5.72%	4.22%
Age 6-7 years	11.12%	12.19%	13.32%	9.69%	10.47%	15.68%	12.54%
Age 8-9 years	9.70%	14.61%	15.14%	10.24%	9.19%	16.99%	14.21%
Age 10-11 years	8.75%	14.04%	15.84%	10.60%	8.74%	16.41%	15.18%
Age 12-14 years	11.23%	19.49%	23.70%	16.20%	11.87%	21.40%	24.05%
Age 15-18 years	12.99%	22.47%	28.37%	22.12%	14.73%	23.79%	29.81%
Age 19-20 years	4.22%	N/A	N/A	11.31%	5.47%	N/A	N/A
Geographic Location							
Urban	83.20%	84.46%	92.08%	96.70%	91.47%	92.10%	92.11%
Rural	15.56%	14.45%	5.07%	3.17%	7.30%	5.00%	5.19%
Missing	1.24%	1.08%	2.85%	0.13%	1.23%	2.89%	2.70%
Race and Ethnicity							
Non-Hispanic White	18.21%	N/A	N/A	N/A	29.89%	N/A	N/A
Non-Hispanic Black	15.45%	N/A	N/A	N/A	29.39%	N/A	N/A
Hispanic	59.42%	N/A	N/A	N/A	29.65%	N/A	N/A
Other & Unknown	6.92%	N/A	N/A	N/A	11.06%	N/A	N/A

Table 1.6D, Patient Characteristics, 10-14 Years Old (Age Range Targeted by Measure), 2010

	Descriptive Characteristics of Individuals 10-14 Years Enrolled at Least One Month, CY 2010						
	Program 1	Program 2	Program 3	Program 4	Program	Plan 1	Plan 2
Total Number Patients	678,393	263,541	124,914	51,634	426,206	29,214	45,652
Age Group Distribution							
Age 10-11 years	43.79%	41.87%	40.06%	39.57%	42.41%	43.40%	38.69%
Age 12-14 years	56.21%	58.13%	59.94%	60.43%	57.59%	56.60%	61.31%
Geographic Location							
Urban	83.37%	84.41%	92.72%	96.64%	91.40%	92.59%	92.80%
Rural	15.78%	14.50%	5.11%	3.26%	7.33%	5.18%	5.24%
Missing	0.86%	1.09%	2.18%	0.10%	1.27%	2.24%	1.97%
Race and Ethnicity							
Non-Hispanic White	17.56%	N/A	N/A	N/A	30.81%	N/A	N/A
Non-Hispanic Black	16.62%	N/A	N/A	N/A	29.98%	N/A	N/A
Hispanic	59.19%	N/A	N/A	N/A	29.01%	N/A	N/A
Other & Unknown	6.63%	N/A	N/A	N/A	10.20%	N/A	N/A

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) the evidence regarding the effectiveness of sealants in caries prevention, (2) an environmental scan, 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. 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

Dental Sealants for 10-14 Year-Old Children at Elevated Caries Risk measures the percentage of children ages 10-14 years at moderate to high risk for dental caries who received a sealant on a permanent second molar tooth during the reporting year. 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) 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

# Requested	# Received	%Received
600	414	69%

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 sealant procedure and tooth number 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 - Sealant and Tooth Number Codes

Data Extraction. For the second component of validation, assessing whether the specific preventive service of sealant placement and associated tooth type coding are accurately captured by claims data, chart abstraction forms were developed by the research team. 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 sealants applied to the teeth during the visit. If there was evidence of sealant placement, the RHITs then recorded whether sealants were applied to the child's permanent first molar, permanent second molar, and/or "other" tooth type. If there was no indication of the tooth to which the sealant was applied, the tooth number field was coded as "indeterminate." 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 code D1351 (sealants); and, when D1351 was present, recording the associated tooth number (or noted as missing). Permanent first molars were identified in the claims data as tooth numbers 3, 14, 19, and 30; permanent second molars were identified as tooth numbers 2, 15, 18, and 31. 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 - RELEVANCE OF TOOTH TYPE

Evidence-based recommendations advise that sealants be placed on pits and fissures of children's primary and permanent teeth when the tooth, or patient, is at caries risk, with stronger evidence for effectiveness in permanent molars (Beauchamp et al. 2008). Sealants benefit children across a wide age range; however, for greatest effectiveness in caries prevention, it is recommended that sealants be placed on teeth soon after they erupt (US DHHS 2010; CDC 2013). Thus, we also sought to evaluate how well the specifications addressed both the tooth type on which sealants are placed and the timeliness of care provision. The research team ran frequency distributions of sealant placement by tooth number and age range for three programs. Specifically, the percentage of children with (1) any sealants (regardless of tooth type), (2) sealants on permanent first

molars, and (3) sealants on permanent second molars was assessed by age for children enrolled at least one month in the program.

Citations

Beauchamp J, Caufield PW, Crall JJ, Donly K, Feigal R, Gooch B, et al. Evidence-based clinical recommendations for the use of pit-and-fissure sealants: a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 2008;139(3):257-268.

Centers for Disease Control and Prevention. 2013. Dental Sealants. Available at: http://www.cdc.gov/OralHealth/publications/faqs/sealants.htm. Accessed January 20, 2014.

U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research. Oral health in America : a report of the Surgeon General. Rockville, Md.: U.S. Public Health Service, Dept. of Health and Human Services; 2000.

4. 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 preventive care 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. 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 preventive care 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

	Percentage of All Enrolled Members Included in Different Denominator Definitions				
	Program 1	Program 2	Program 3	Program 4	Program 5
At least 1 month	100%	100%	100%	100%	100%
At least 3 months	95%	85%	84%	93%	94%
At least 6 months	83%	63%	65%	81%	81%
11-12 months	64%	33%	42%	63%	59%

5. ADDITIONAL VALIDITY TESTING - IDENTIFYING ELEVATED RISK WITH CLAIMS DATA

Evidence-based guidelines indicate that sealants are most effective for children at higher risk for caries (see Measure Evidence Form). Thus, inclusion in the denominator is limited to children identified as being at moderate to high risk for caries. Administrative claims data for dental claims typically do not include diagnostic codes. Procedure codes for risk assessment that identify moderate and high risk were included in the measure logic. However, because these are newer codes, additional logic was included to identify children with recent history of restorations, which are indicative of caries. A systematic review found that prior caries experience to be an important predictor of future risk (Zero D, Fontana M, Lennon AM. 2001. Clinical applications and outcomes of using indicators of risk in caries management. J Dent Educ. 2001 Oct;65(10):1126-32.) Expert consensus and validation through chart reviews was done to finalize the procedure codes (indicated in the measure specifications) used to identify elevated risk. The test data results reported in

this application demonstrate that it is feasible to use these validated codes to identify children at elevated risk who should receive preventive services.

6. 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

Number of Procedure Codes	Record and Procedure Code on Claim Correlate	Record Did Not Correlate with Procedure Code on Claim	Unable to Determine Correlation
1,135	94.04%	4.22%	1.75%

B. Critical Data Element Validation - Sealant and Tooth Number Codes
To assess whether the specific preventive service of dental sealants and associated tooth type 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 for sealants and tooth number. Agreement (concordance) for sealant placement was 95%. Sensitivity of sealant placements was moderately high (77.8%) and specificity was very high (98.8%). Sensitivity was not as strong for second permanent molars (50%), but specificity was very high (100%). The positive predictive and negative predictive values were both high (>93%) for sealant placement with a lower negative predictive value for the specific tooth type. 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 for sealants was also very high at 0.8205 indicating "almost perfect" agreement. For dates of service in which there was agreement with the administrative data that sealants had been applied (n=84), we then assessed whether there was agreement on tooth type using the following categories: permanent first molar, permanent second molar, and other teeth. We report here on the findings for permanent second molar which is the focus of the proposed measure (we had similar findings for first molars). Overall, the simple agreement percentage was 88% for permanent second molars. The corresponding kappa statistic value was 0.604, indicating "substantial" agreement.

Table 2b2.3-2 Agreement between Record and Administrative Data for Specific Services

	Concordance	Prevalence	Sensitivity	Specificity	PPV	NPV	Kappa
Sealants Applied	95.22%	0.172	0.778	0.988	0.933	0.955	0.820
Dates of service: 613			(0.686-0.850)	(0.974-0.995)	(0.855-0.973)	(0.933-0.971)	(0.758-0.882)
# indeterminate: 4							
Second Molar (if sealant)	88.10%	0.238	0.500	1.000	1.000	0.866	0.604
Dates of service: 613			(0.279-0.722)	(0.930-1.000)	(0.656-1.000)	(0.761-0.930)	(0.392 - 0.815)
# indeterminate: 0							

95% confidence intervals indicated in parentheses

Our findings are similar 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. For sealants, they found lower sensitivity (73%), higher specificity (100%) and similar kappa value (0.84) of billing data compared to direct observation. (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

Sealants on a Permanent Molar Tooth was identified through the Delphi rating process as a high-scoring measure concept with a mean importance score of 7, mean feasibility score of 8, and mean validity score of 7, 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.] Thus, the measure has face validity. However, gaps were identified with existing measures, including not associating tooth type and age range, lack of clear specifications, and lack of standardization. The proposed measure overcomes these limitations.

Content Validity. In addition, the measure also demonstrates **content validity** – the extent to which the measure specifications reflect the intended domain of care. This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (sealants) as well as the specific tooth type for which the evidence is the strongest (permanent molar) and the timing of sealant placement to maximize effectiveness (shortly after eruption – 10-14 years of age for permanent second molars). Please see the Measure Evidence Form for more details.

3. ADDITIONAL VALIDITY TESTING - RELEVANCE OF TOOTH NUMBER

Analysis of sealant placement by tooth type and age range validated the importance of including specific teeth numbers in the measure specifications to identify permanent first molars and permanent second molars and associating those tooth numbers with the corresponding appropriate age ranges (6-9 years and 10-14 years, respectively) in order to have reliable indicators of whether children are getting recommended and timely prevention. Table 2b2.3-3 indicates the percentage of children in each of three programs who had (1) a sealant placed on any tooth, (2) a sealant placed on a permanent first molar, and (3) a sealant placed on a permanent second molar; the same child could be included in more than one category. In all programs, the percentage of children with "any sealants" is greater than the percentage of children with sealants specifically on permanent second molars. Children in this age group may receive sealants or replacement sealants on premolars or permanent first molars, which confounds findings about whether permanent second molars are being sealed

when the tooth type is not identified in the measure specifications. The differences between the percentage of children with "any sealants" and those with sealants on permanent second molars are most dramatic for Program 1 compared to Programs 3 and 4 due to differences in benefit coverage between the programs; Program 1 did not condition reimbursement for sealants on tooth type. These results indicate that children ages 10-14 years may have teeth other than permanent second molars sealed that would get captured in the numerator and inflate the measure score if the type of tooth is not specified, resulting in misleading comparisons of performance between programs. Thus, the research team concluded that the incorporation of teeth numbers in the DQA specifications is a significant and important improvement over existing sealant measures that have lacked this specificity.

Table 2b2.3-3 Sealant Placement by Age and Tooth Type

	Program 1			Program 3			Program 4		
Age (years)	% with Any Sealants (Any Tooth)	% with Sealant on Permanent 1st Molars	% with Sealant on Permanent 2nd Molars	% with Any Sealants (Any Tooth)	% with Sealant on Permanent 1st Molars	% with Sealant on Permanent 2nd Molars	% with Any Sealants (Any Tooth)	% with Sealant on Permanent 1st Molars	% with Sealant on Permanent 2nd Molars
6	25.02%	13.73%	0.04%	6.42%	6.32%	0.04%	8.21%	7.58%	0.01%
7	34.44%	26.20%	0.06%	15.03%	14.95%	0.09%	21.21%	20.92%	0.09%
8	31.02%	21.56%	0.08%	15.52%	15.49%	0.15%	18.85%	18.70%	0.12%
9	29.80%	14.00%	0.28%	12.45%	12.34%	0.18%	11.35%	11.06%	0.19%
10	35.36%	9.91%	1.87%	10.36%	9.90%	0.88%	7.63%	6.77%	0.74%
11	40.45%	7.42%	6.92%	10.18%	8.78%	3.07%	7.70%	4.92%	3.18%
12	40.96%	5.36%	12.76%	10.46%	7.67%	6.29%	11.99%	4.57%	9.05%
13	36.20%	3.73%	14.40%	10.40%	6.89%	8.27%	14.94%	4.04%	13.34%
14	29.85%	2.82%	11.64%	9.07%	5.93%	8.08%	12.44%	3.32%	11.51%

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?*)
As noted above, the overall agreement between the administrative claims data and dental record data was high based on both simple agreement and using the more conservative Kappa statistic. Although the agreement for the specific tooth type was not as strong as for sealant application in general, it was still "substantial," and we believe that data concordance will improve with increasing accountability as is often the case when new performance measures are implemented. Overall, we interpret these findings as evidence that validates the accuracy of administrative claims data for performance measurement purposes. These empirical findings, combined with our face validity assessments of the measure score, lead us to conclude that both the data elements and the measure score represent valid measures of sealant placement prevalence among 10-14 year olds. In addition, our testing indicated that the incorporation of tooth number as part of the measure specifications was important for ensuring that the measure captures sealant placement on the tooth type (permanent second molars) for which there is the strongest evidence of effectiveness among this age group.

2b3. EXCLUSIONS ANALYSIS

NA X☐ no exclusions — *skip to section 2b4*

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.
Not applicable.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with risk factors**
☐ **Stratification by risk categories**
☐ **Other,**

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*)
Not applicable.

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

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 had a sealant, the dichotomous outcome of had/did not have a sealant 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/Plan, Year, Measure Score as % (Measure Score, SD, Lower 95% CI, Upper 95% CI)

Program 1, CY 2011:	11.08%	(	0.1108 ,	0.0005 ,	0.1099 ,	0.1117)
Program 2, CY 2011:	10.59%	(	0.1059 ,	0.0010 ,	0.1039 ,	0.1079)
Program 3, CY 2011:	10.58%	(	0.1058 ,	0.0018 ,	0.1022 ,	0.1094)
Program 4, CY 2011:	10.84%	(	0.1084 ,	0.0026 ,	0.1034 ,	0.1134)
Program 1, CY 2010:	10.48%	(	0.1048 ,	0.0005 ,	0.1038 ,	0.1058)
Program 2, CY 2010:	7.67%	(	0.0767 ,	0.0009 ,	0.0749 ,	0.0785)
Program 3, CY 2010:	10.36%	(	0.1036 ,	0.0018 ,	0.1000 ,	0.1072)
Program 4, CY 2010:	12.70%	(	0.1270 ,	0.0027 ,	0.1217 ,	0.1323)
Program 5, CY 2010:	8.44%	(	0.0844 ,	0.0011 ,	0.0823 ,	0.0865)
Plan 1, CY 2011:	9.64%	(	0.0964 ,	0.0027 ,	0.0911 ,	0.1017)
Plan 2, CY 2011:	11.06%	(	0.1106 ,	0.0025 ,	0.1056 ,	0.1156)
Plan 1, CY 2010:	10.10%	(	0.1010 ,	0.0045 ,	0.0923 ,	0.1097)
Plan 2, CY 2010 :	10.05%	(	0.1005 ,	0.0032 ,	0.0943 ,	0.1067)

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 (both years) and the plans (one year) (Table 2b5.2).

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

	Chi-Square Value	*p* - value
Program Results, 2011	22.70	<0.0001
Program Results, 2010	899.80	<0.0001
Plan Results, 2011	14.46	0.0001
Plan Results, 2010	0.01	0.9203

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 the program and plan levels. At the plan level, statistically significant differences were detected in 2011, but not in 2010. This is consistent with a greater difference in performance between the two plans in 2011 (9.64% and 11.06%) than in 2010 when the rates were almost equal (10.10% and 10.05%). This is precisely the purpose of performance measurement - to detect when there are differences in performance. In 2010, there was no appreciable difference in performance between the two plans. Collectively, however, it is clear that this measure detects differences in performance on the measure scores when they do exist. Our findings are consistent with evidence reported earlier in this application documenting disparities in sealant receipt among children. Thus, this measure informs performance improvement efforts by allowing plans and programs to identify and monitor performance gaps and disparities in performance 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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

This measure is specified for reporting at the program and plan level and there are no current plans for developing an eMeasure (eCQM) at these levels.

Our understanding is that the Feasibility Score Card is only for eMeasures; consequently, we have not submitted this. Feasibility criteria were met during the initial endorsement review.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

This measure relies on standard data elements in administrative claims data (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes. A key advantage of using administrative claims data is that the time and cost of data collection for performance measurement purposes are relatively low because these data are already collected for other purposes.

Initial feasibility assessments were conducted using the RAND-UCLA modified Delphi process to rate the measure concepts with feasibility as one component of the assessment. On a 1-9 point scale, this measure concept was rated as an 8 or "definitely feasible" by the expert panel. During the empirical testing phase, our testing found that all of the critical data elements except one had missing/invalid data of <1% (Data 3c.1.), meeting or exceeding the guidance from the Centers for Medicare and Medicaid Services regarding acceptable error rates. The exception was tooth number associated with sealant procedure codes.

Missing/invalid data rates ranged from 0.15% to 15%, with most programs having missing/invalid rates <5%. We do not view the higher rates among a subset of the programs as a threat to feasibility, however. The high compliance by the majority of programs indicates that it is feasible to obtain missing and invalid rates of <1%. The Centers for Medicare and Medicaid Services already requires state Medicaid programs to report sealants placed on permanent molars among enrolled children, which requires data on tooth number, and tooth number also is typically required for reimbursement. During measure development and testing, the measure specifications were made available through a publicly accessible website for public comment with additional broad email dissemination to a wide range of stakeholders. No concerns regarding feasibility of collecting any of the data elements were raised during this process.

Citation: Centers for Medicare & Medicaid Services. Medicaid and CHIP Statistical Information System (MSIS) File Specifications and Data Dictionary. 2010; http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MSIS/downloads/msisdd2010.pdf. Accessed August 10, 2013.

Data 3c.1 Percentage of Missing and Invalid Values for Critical Data Elements

PROGRAM 1
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Tooth number: 6.18%
Date of Service: 0.01%
Rendering Provider ID: 0.28%

PROGRAM 2
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Tooth number: 15.31%
Date of Service: 0.00%
Rendering Provider ID: 0.00%

PROGRAM 3
Member ID: 0.27%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.28%
Tooth number: 0.18%
Date of Service: 0.00%
Rendering Provider ID: 0.18%

PROGRAM 4
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.01%
Tooth number: 2.47%
Date of Service: 0.00%
Rendering Provider ID: 0.61%

PROGRAM 5
Member ID: 0.43%
Date of Birth: 0.02%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Tooth number: 0.15%

Date of Service: 0.00%
Rendering Provider ID: 0.67%

Endorsement Maintenance Update: There have been no reports of feasibility issues with implementing this measure. Please see Use and Usability section.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
This measure is intended to be transparent and available for widespread adoption. As such, it was purposefully designed to avoid using software or other proprietary materials that would require licensing fees. The measure specifications, including a companion User Guide, is accessible through a website and can be used free of charge for non-commercial purposes. The main requirements of users is to ensure the quality of their source data and expertise to program the measures within their information systems, following the clear and detailed specifications. Technical assistance is available to users.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting Texas Health and Human Services Commission: Texas Medicaid/CHIP https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf Payment Program Texas Health and Human Services Commission: Texas Medicaid/CHIP https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf Quality Improvement (external benchmarking to organizations) Covered California http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf Michigan Healthy Kids Dental https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids Quality Improvement (Internal to the specific organization) State Medicaid Agencies http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

4a1.1 **For each CURRENT use, checked above (update for maintenance of endorsement), provide:**

- Name of program and sponsor

- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

1. Program and Sponsor: Texas Health and Human Services Commission: Texas Medicaid/CHIP

https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf

Purpose: Payment Program and Public Reporting

This measure has been adopted by the Texas Health and Human Services Commission as part of the Texas CHIP and Medicaid Dental Services Pay-for-Quality (P4Q) program. [Texas HHSC Uniform Managed Care Manual, Chapters 6.2.15. Effective Date 09/01/2017, Version 2.0].

This measure was also present in earlier iterations of the Texas Medicaid and CHIP quality programs since initial endorsement. We are referencing current use for this update.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies to the state of Texas CHIP and Medicaid programs (statewide application). There are two dental plans (i.e., the accountable entities) that serve Texas CHIP and Medicaid. In June 2017, there were 3,359,770 children enrolled in Texas Medicaid and CHIP (https://hhs.texas.gov/about-hhs/records-statistics/data-statistics/healthcare-statistics).

Level of Measurement and Setting: The measure is implemented at the plan and program level within the Texas Medicaid and CHIP programs.

2. Covered California, the California Health Benefit Exchange

http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf
http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-QDP-Issuer-Contract-and-Attachments.pdf

Purpose: Quality Improvement

This measure is included in the Covered California Qualified Health Plan Issuer Contract for 2017-019 For the Individual Market and the Covered California Qualified Dental Plan Issuer Contract for 2017-2019. The measure is to be reported annually.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies statewide. In March 2017 there were 85,000 enrollees 0-18 years old in CC health plans (which may offer dental benefits and would therefore report on the dental quality measures). There were 5,100 children enrolled specifically in Qualified Dental Plans. (http://hbex.coveredca.com/data-research/)

Level of Measurement and Setting. The measure is implemented at the plan level with the Covered California program.

3. State Medicaid Agencies

http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

(Note: To access the data, a public user account must be created. We can help facilitate access to the data if needed.)

Purpose: Quality Improvement

The Medicaid|Medicare|CHIP Services Dental Association conducts an annual survey of state Medicaid programs and collects data specifically on which programs report Dental Quality Alliance measures.

In its 2015 profile (the most recent available), 11 states reported that they currently use this measure in the Medicaid and/or CHIP programs.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The 11 states are: Alabama, Colorado, Connecticut, Florida, Idaho, Illinois, Nevada, Oklahoma, Rhode Island, South Carolina,and West Virginia. Data are not provided on the number of accountable entities included.

4. Michigan Healthy Kids Dental Program
https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids

Note: Select Schedule A Work Statement link under File Attachments

Purpose: Quality Improvement

The Michigan Healthy Kids Dental Program has included this measure in the set of measures included in its Performance Monitoring Standards, which is currently included in the Request for Proposals and will be included in the contracts between the contracted dental plans and the State of Michigan.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The Healthy Kids dental program covers children enrolled in Michigan's Medicaid program statewide. The state intends to award two contracts. There are approximately 955,000 enrollees served by the Healthy Kids Dental Program.

Additional Information:
This measure was one of ten performance measures that focused on Dental Caries Prevention and Disease Management among children and were approved by the DQA. 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)

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not applicable.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Per the annual survey conducted by the Medicaid|Medicare|CHIP Services Dental Association (MSDA), 11 Medicaid/CHIP agencies are implementing this measure. The measure is part of measure set included in the Request for Proposal (RFP) released by the Michigan Healthy Kids Dental Program. This measure is included in the Pay-For-Quality program and is publicly reported for Texas Medicaid/CHIP. Additionally, this measure is a requirement for the Qualified Dental Plans to report to the Covered California, the state-based marketplace in California.

The DQA provides technical assistance to these and other users of DQA measures through webinars, resource document development, and one-on-one staff support. The DQA has an Implementation Committee dedicated to developing implementation and improvement resources.

In order to ensure transparency, incorporate learnings from implementation, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting. Implementation has mostly focused on addressing questions related to how to use the measures in the context of broader quality improvement and clarifying questions related to the specifications.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**

In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews. The DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**

Describe how feedback was obtained.

In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews. The DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.2. **Summarize the feedback obtained from those being measured.**

A dental benefits administrator (DBA) has suggested that the DQA consider adding patient exclusions to the measure. The DQA considered exclusions previously during initial measure development and during annual reviews. Exclusions were not included due to concerns about the introduction of biased measurement, increasing measurement complexity, and adversely affecting

implementation feasibility. However, the DQA continues to monitor this issue and will revisit it during the 2018 annual review. The DQA has invited the DBA to present its suggestion with supporting data to the DQA. The DQA has also invited other DBAs and Medicaid program administrators to provide input. All of this stakeholder feedback will be incorporated into the next annual review.

4a2.2.3. **Summarize the feedback obtained from other users**
No additional feedback.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
The DQA considered exclusions during initial measure development and during annual reviews. Exclusions were not included due to concerns about the introduction of biased measurement, increasing measurement complexity, and adversely affecting implementation feasibility. However, the DQA continues to monitor this issue and will revisit it during the 2018 annual review. The DQA has invited the DBA to present its suggestion with supporting data to the DQA. The DQA has also invited other DBAs and Medicaid program administrators to provide input. All of this stakeholder feedback will be incorporated into the next annual review.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures either have only limited baseline scores or will start reporting measures within the next year.

We are only aware of repeat measurements within the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after it was approved by the Dental Quality Alliance and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 475976, 16.78, 17.10, 16.59
2015, 527493, 16.63, 16.48, 16.90

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 102148, 12.59, 14.08, 12.96
2015, 70216, 12.59, 13.90, 14.28

These data suggest fairly stable rates over the two-year period. However, as noted above, these are initial performance data for one program; additional time may be needed to see improvement within this program. Most measure users are just now getting their quality measurement programs underway.

4b2. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**

No unintended or negative consequences have been identified.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization of Related Measures
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable.

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Appendix_Sealants1014.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Dental Association on behalf of the Dental Quality Alliance
Co.2 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-
Co.3 **Measure Developer if different from Measure Steward:** American Dental Association on behalf of the Dental Quality Alliance
Co.4 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This project is headed by the DQA through its Measure Development and Maintenance Committee (formerly Research and Development Committee). The following individuals were responsible for executing and overseeing all scientific aspects of this project.

- Craig W. Amundson, DDS, General Dentist, HealthPartners, National Association of Dental Plans. Dr. Amundson serves as chair for the Committee.
- Mark Casey, DDS, MPH, Dental Director, North Carolina Department of Health and Human Services Division of Medical Assistance
- Natalia Chalmers, DDS, PhD, Diplomate, American Board of Pediatric Dentistry, Director, Analytics and Publication, DentaQuest Institute
- Frederick Eichmiller, DDS, Vice President & Science Officer, Delta Dental of Wisconsin
- Chris Farrell, RDH, BSDH, MPA, Oral Health Program Director, Michigan Department of Health and Human Services

This group oversees the maintenance process of the measures. All work of this Committee was distributed for review and formal vote and approval by the entire Dental Quality Alliance. (http://ada.org/dqa) The DQA is made up of representatives from 38 stakeholder organizations.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2013
Ad.3 **Month and Year of most recent revision:** 01, 2017
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2018

Ad.7 **Disclaimers:** Dental Quality Alliance measures and related data specifications, developed by the Dental Quality Alliance (DQA), are intended to facilitate quality improvement activities. These Measures are intended to assist stakeholders in enhancing quality of care. These performance Measures are not clinical guidelines and do not establish a standard of care. The DQA has not tested its Measures for all potential applications.
Measures are subject to review and may be revised or rescinded at any time by the DQA. The Measures may not be altered without the prior written approval of the DQA. The DQA shall be acknowledged as the measure steward in any and all references to the measure.
Measures developed by the DQA, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and DQA. Neither the DQA nor its members shall be responsible for any use of these Measures.
THE MEASURES ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND
Limited proprietary coding is contained in the Measure specifications for convenience.
For Proprietary Codes:

Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. The DQA, American Dental Association (ADA), and its members disclaim all liability for use or accuracy of any terminologies or other coding contained in the specifications. THE SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.
Ad.8 **Additional Information/Comments:** In 2008, the Centers for Medicare and Medicaid Services (CMS) asked the ADA to lead the development of a broad coalition of organizations that would lead dentistry to improve the oral health of Americans through quality measurement and quality improvement. The ADA subsequently established the DQA. The DQA is a multi-stakeholder alliance comprised of approximately 38 stakeholders (with organizations as members) from across the oral health community, including federal agencies, third-party payers, professional associations, and an individual member from the general public. The DQA's mission is to advance the field of performance measurement to improve oral health, patient care, and safety through a consensus building process.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2511
Measure Title: Utilization of Services, Dental Services
Measure Steward: American Dental Association on behalf of the Dental Quality Alliance
Brief Description of Measure: Percentage of enrolled children under age 21 years who received at least one dental service within the reporting year.
Developer Rationale: Inequalities in oral health status and inadequate use of oral health care services are well documented (Dye, Li, and Thorton-Evans 2012; IOM 2011a, 2011b; US DHHS 2010). Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, L i, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009).

Improving access to care through the oral health care delivery system is critical to improving oral health outcomes and addressing oral health disparities. In the IOM report, Improving Access to Oral Health Care for Vulnerable and Underserved Populations, there were four overall conclusions. The first conclusion was: "Improving access to oral health care is a critical and necessary first step to improving oral health outcomes and reducing disparities" (IOM 2011b). Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions. However, there are significant performance gaps and disparities in access. Untreated dental caries occurs among 25% of children living in poverty compared with 10.5% of children living above poverty (Dye, L i, and Thorton-Evans 2012). Approximately 75% of children younger than age 6 years did not have at least one visit to a dentist in the previous year (Edlestein 2009). Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), there are significant variations in use of dental services across states, ranging from approximately 25% to 69% (CMS-416 data, FY 2011). Even among the highest performing states, more than one-fourth of publicly-insured children do not have a dental visit during the year. Similar variation between states is observed among children 0-20 years of age enrolled in commercial dental plans (ADA 2013).

The proposed measure, Utilization of Services – Dental Services, captures whether a child received any dental services during the year and, therefore, also measures access to oral health care – the "critical and necessary first step to improving oral health outcomes and reducing disparities" (IOM 2011b). This measure also includes important stratifications by the children's age. Utilization of Services allows plans and programs to identify the effectiveness of efforts in improving access to oral health services and target performance improvement initiatives accordingly.

This measure is a program/plan specific measure that contributes to the Healthy People 2020 Objective OH 7 that calls for increasing the proportion of children, adolescents, and adults who used the oral health care system in the past year. This is a leading health indicator.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information,

evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

Numerator Statement: Unduplicated number of children under age 21 years who received at least one dental service
Denominator Statement: Unduplicated number of enrolled children under age 21 years
Denominator Exclusions: Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Sep 18, 2014 **Most Recent Endorsement Date:** Sep 18, 2014

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- Dental decay causes significant short- and long-term adverse consequences for children's health and functioning. Moreover, there are documented disparities in untreated dental caries and receipt of dental services (see sections 1b4 and 1b5). The proposed measure, Utilization of Services – Dental Services, captures whether a child received any dental services during the year and, therefore, also measures access to oral health care. The American Academy of Pediatric Dentistry (AAPD) recommends that all children have a dental home established by 12 months of age, which it defines as "the ongoing relationship between the dentist and the patient, inclusive of all aspects of oral health care delivered in a comprehensive, continuously accessible, coordinated, and family-centered way".
    - NICE Guidelines: Although NICE has a detailed method for grading evidence in developing clinical guidelines, the report does not contain the specific grades assigned for the evidence associated with each clinical guideline.
    - AAPD Guidelines: Evidence grades were not assigned.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☒ **The developer provided updated evidence for this measure:**

Updates:

- A more recent Cochrane review evaluated this topic (Riley et al. 2013). The Cochrane review only included randomized controlled trials; thus, only 1 study was included. The main finding of that study was: "For three to five-year olds with primary teeth, the mean difference (MD) in dmfs increment was -0.90 (95% CI -1.96 to 0.16) in favour of 12-month recall. For 16 to 20-year olds with permanent teeth, the MD in DMFS increment was -0.86 (95% CI -1.75 to 0.03) also in favour of 12-month recall."

Citation:
Riley P, Worthington HV, Clarkson JE, Beirne PV. Recall intervals for oral health in primary care patients. Cochrane Database of Systematic Reviews 2013, Issue 12.

Question for Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review, but does note a recent Cochrane review collated all evidence and reached the same conclusions that supported the original guideline. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm
Process measure based on systematic review (Box 3)→ Empirical evidence submitted (Box 7)→ Empirical evidence includes all studies in body of evidence (Box 8) → Rate as Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The Developer used data from five sources and refers to "program" level information and "plan" level information (Texas Medicaid, Texas CHIP, Florida CHIP, and Florida Medicaid programs, as well as national commercial data from Dental Service of Massachusetts, Inc.). The developer presented the total number of children enrolled in each program/plan. In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011.
- The data source and sample size is sufficient to assess gaps in performance. The performance range of **28% to 74**% in CY 2010 (year in which data were available for all five programs) indicates a significant performance gap overall. Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), there are significant variations in use of dental services across states, ranging from approximately 25% to 69% (CMS-416 data, FY 2011).
- The developer did not provide more recent performance data, stating that due to the start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting.

Disparities

- The developer's findings demonstrate that there are disparities by age, geographic location (all except one program), and race/ethnicity. It also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), Medicaid program type, commercial product line, and preferred language for program communications. The developer detected disparities for each of these factors, but data on all of these characteristics were not consistently available

for all programs, so it presented disparities data on those characteristics most consistently available and with the greatest standardization (i.e. race/ethnicity and geographic location).

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Staff Scientific Acceptability Logic

*The original testing was submit as permitted by NQF.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure relies on **standard data elements in administrative claims data** (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes.

Update: The developer states there have been no reports of feasibility issues with implementing this measure.

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- Texas Health and Human Services Commission: Texas Medicaid/CHIP Pay-for-Quality (P4Q) program. https://hhs.texas.gov/sites/default/files//documents/lawsregulations/ handbooks/umcm/6-2-15.pdf

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- In 2016, the Dental Quality Alliance (DQA) expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from six state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from eight dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results
The developer notes that it is only aware of repeat measurements within the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after it was approved by the DQA and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 2698361, 69.61, 71.02, 68.28
2015, 2929975, 71.49, 72.70, 69.97

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 452976, 61.96, 64.62, 61.67
2015, 341937, 65.90, 70.44, 67.36

These data suggest a trend in improvement over time. However, as noted above, these are initial performance data for one program. Most measure users are just now getting their quality measurement programs underway.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- No unintended or negative consequences were identified by the developer.

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- N/A

Harmonization

- N/A

Committee pre-evaluation comments
Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Staff Scientific Acceptability Logic

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☒Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

   ☐Yes (go to Question #5)

   ☒No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☐Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
   Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☐High (go to Question #8)
   ☐Moderate (go to Question #8)
   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☒Yes (go to Question #9)
   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☒Yes (go to Question #10)
   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

   ☒Yes (go to Question #2)

   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

   ☐Yes (please explain below then go to Question #3)

   ☐No (go to Question #3)

   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?

   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

   ☐Yes (please explain below then go to Question #4)

   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

   ☐Yes (please explain below then go to Question #5)

   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

   ☐Yes (please explain below then go to Question #6)

   ☒No (go to Question #6)

☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☐Yes (go to Question #11)
    ☒No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☐Yes (go to Question #12)

☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
    ☐High (go to Question #14)
    ☐Moderate (go to Question #14)
    ☐Low (please explain below then go to Question #13)
    ☐Insufficient

13. Was other validity testing reported?
    ☒Yes (go to Question #14)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?
    TIPS: Prior validity studies of the same data elements may be submitted
    ☒Yes (go to Question #15)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*
    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
    ☒Yes (go to Question #16)
    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
    ☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
    ☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Utilization of Services, Dental Services

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 2/10/2014

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☐ Health outcome:

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome:

X☐ Process: Receipt of dental services during the reporting period

☐ Structure:

☐ Other:

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not applicable.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

Not applicable.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

As described in 1b1 (Importance), dental caries is the most common chronic disease in children in the U.S., and a significant percentage of children have untreated dental caries. Dental decay causes significant short- and long-term adverse consequences for children's health and functioning. Moreover, there are documented disparities in untreated dental caries and receipt of dental services (see sections 1b4 and 1b5). The proposed

measure, Utilization of Services – Dental Services, captures whether a child received any dental services during the year and, therefore, also measures access to oral health care. The Institute of Medicine has identified improving access to oral health care as a "critical and necessary first step to improving oral health outcomes and reducing disparities."

Institute of Medicine and National Research Council. Improving access to oral health care for vulnerable and underserved populations. Washington, D.C.: National Academies Press; 2011.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

American Academy of Pediatric Dentistry. 2013. "Guideline on Periodicity of Examination, Preventive Dental Services, Anticipatory Guidance/Counseling, and Oral Treatment for Infants, Children, and Adolescents. " Available at: http://www.aapd.org/media/Policies_Guidelines/G_Periodicity.pdf.

American Academy of Pediatric Dentistry. 2012. "Policy on the Dental Home. " Available at: http://www.aapd.org/media/Policies_Guidelines/P_DentalHome.pdf.

American Academy of Pediatrics Section on Pediatric Dentistry and Oral Health. 2008. "Policy Statement: Preventive Oral Health Intervention for Pediatricians." Pediatrics 122(6): 1387-94. Available at: http://pediatrics.aappublications.org/content/122/6/1387.full.

National Institute for Health and Care Excellence (NICE). 2004. Clinical Guidelines. "CG19: Dental Recall – Recall Interval between Routine Dental Examinations." Available at: http://guidance.nice.org.uk/CG19.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

The American Academy of Pediatric Dentistry (AAPD) recommends that all children have a dental home established by 12 months of age, which it defines as "the ongoing relationship between the dentist and the patient, inclusive of all aspects of oral health care delivered in a comprehensive, continuously accessible, coordinated, and family-centered way" (AAPD, Dental Home Definition, http://www.aapd.org/media/Policies_Guidelines/P_DentalHome.pdf). Consistent with the dental home concept, national guidelines from the American Academy of Pediatric Dentistry (AAPD) and the American Academy of

Pediatrics (AAP) recommend that children receive oral health services by 1 year of age and have regular visits thereafter. The most common recall interval is six months. However, evidence-based guidelines indicate that the recall schedule should be tailored to individual needs based on assessments of existing disease and risk of disease (e.g., caries risk) with a recommended recall frequency for routine visits ranging from 3 months to no more than 12 months for individuals younger than 18 years of age (NICE 2004).

Age of First Visit

"The first examination is recommended at the time of the eruption of the first tooth and no later than 12 months of age." (p. 114 of AAPD Clinical Guidelines).

"Children who have a dental home are more likely to receive appropriate preventive and routine oral health care. Referral by the primary care physician or health provider has been recommended, based on risk assessment, as early as six months of age, six months after the first tooth erupts, and no later than 12 months of age."

"Every child should have a dental home established by 1 year of age." (American Academy of Pediatrics Section on Pediatric Dentistry and Oral Health. 2008. "Policy Statement: Preventive Oral Health Intervention for Pediatricians." Pediatrics 122(6): 1387-94; at page 1391).

Supporting evidence cited in AAPD Guidelines and Policy:

American Academy of Pediatric Dentistry. Policy on the dental home. Pediatr Dent 2012;34(special issue):24-5.

American Academy of Pediatrics. Oral health risk assessment timing and establishment of the dental home. Pediatr 2003:11(5):1113-6. Reaffirmed 2009;124(2):

Berg JH, Stapleton FB. Physician and dentist: New initiatives to jointly mitigate early childhood oral disease. Clin Pediatr 2012:51(6):531-7.

Nowak AJ, Casamassimo PS. The dental home: A primary oral health concept. J Am Dent Assoc 2002;133 (1):93-8.

Nowak AJ. Rationale for the timing of the first oral evaluation. Pediatr Dent 1997;19(1):8-11.

Recall Interval

"The recommended interval between oral health reviews should be determined specifically for each patient and tailored to meet his or her needs, on the basis of an assessment of disease levels and risk of or from dental disease." (NICE Guidelines, 2004, p. 40)

"The shortest interval between oral health reviews for all patients should be 3 months." (NICE Guidelines, 2004, p. 41) Note: NICE uses the term "oral health reviews"

"The longest interval between oral health reviews for patients younger than 18 years should be 12 months." (NICE Guidelines, 2004, p. 41)

- Rationale: "There is evidence that the rate of progression of dental caries can be more rapid in children and adolescents than in older people, and it seems to be faster in primary teeth than in permanent teeth

(see Chapter Three, Section 3.1.2.) Periodic developmental assessment of the dentition is also required in children. Recall intervals of no longer than 12 months give the opportunity for delivering and reinforcing preventive advice and for raising awareness of the importance of good oral health. This is particularly important in young children, to layout the foundations for life-long dental health." (NICE Guidelines, 2004, p. 41)

"For practical reasons, the patient should be assigned a recall interval of 3, 6, 9, or 12 months if he or she is younger than 18 years, or 3, 6, 9, 12, 15, 18, 21, or 24 months if he or she is aged 18 years or older." (NICE Guidelines, 2004, p. 41)

"The most common interval of examination is six months; however, some patients may require examination and preventive services at more or less frequent intervals, based upon historical, clinical, and radiographic findings." (p. 115 of AAPD Clinical Guidelines)

Supporting evidence cited by AAPD Clinical Guidelines:

Beil HA, Rozier RG. Primary health care providers' advice for a dental checkup and dental use in children. Pediatr 2010;126(2):435-41.

Pahel BT, Rozier RG, Stearns SC, Quiñonez RB. Effectiveness of preventive dental treatments by physicians for young Medicaid enrollees. Pediatr 2011;127(3):682-9.

Diangelis AJ, Andreasen JO, Ebeleseder KA, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Dental Injuries: 1. Fractures and luxations of permanent teeth. Dent Traumatol 2012;28(1):2-12.

Andersson L, Andreasen JO, Day P, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Dental Injuries: 2. Avulsion of permanent teeth. Dent Traumatol 2012;28(2):88-96.

Malmgren B, Andreasen JO, Flores MT, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Injuries: 3. Injuries in the primary dentition. Dent Traumatol 2012;28(3):174-82.

Patel S, Bay RC, Glick M. A systematic review of dental recall intervals and incidence of dental caries. J Am Dent Assoc 2010;141(5):527-39.

American Academy of Pediatric Dentistry. Guideline on prescribing dental radiographs. Pediatr Dent 2012;34(special issue):299-301.

American Dental Association Council on Scientific Affairs.The use of dental radiographs; Update and recommendations. J Am Dent Assoc 2006;137(9):1304-12.

Greenwell H, Committee on Research, Science and Therapy American Academy of Periodontology. Guidelines for periodontal therapy. J Periodontol 2001;72(11):1624-8.

Califano JV, Research Science and Therapy CommitteeAmerican Academy of Periodontology. Periodontal diseases of children and adolescents. J Periodontol 2003;74(11):1696-704.

Clerehugh V. Periodontal diseases in children and adoles¬cents. British Dental J 2008;204(8):469-71.845.

Benefits Obtained

"Early detection and management of oral conditions can improve a child's oral health, general health and well-being, and school readiness." (p. 114 of AAPD Clinical Guidelines)

Supporting evidence cited by AAPD Guidelines:

American Academy of Pediatric Dentistry. Policy on early childhood caries: Classifications, consequences, and preventive strategies. Pediatr Dent 2012;34(special issue):50-2.

American Academy of Pediatric Dentistry. Policy on early childhood caries: Unique challenges and treatment options. Pediatr Dent 2012;34(special issue):53-5.

Clarke M, Locker D, Berall G, Pencharz P, Kenny DJ, Judd P. Malnourishment in a population of young children with severe early childhood caries. Pediatr Dent 2006;28(3):254-9.

Dye BA, Shenkin JD, Ogden CL, Marshall TA, Levy SM, Kanellis MJ. The relationship between healthful eating practices and dental caries in children ages 2-5 years in the United States, 1988-1994. J Am Dent Assoc 2004;135(1):55-6.

Jackson SL, Vann WF, Kotch J, Pahel BT, Lee JY. Impact of poor oral health on children's school attendance and performance. Amer J Publ Health 2011;10(10):1900-6.

Every visit provides the opportunity to provide anticipatory guidance, which "is the process of providing practice, developmentally-appropriate information about children's health to prepare parents for the significant physical, emotional, and psychological milestones." (AAPD Clinical Guidelines, p. 116) "Individualized discussion and counseling [anticipatory guidance] should be an integral part of each visit. Topics to be included are oral hygiene and dietary habits, injury prevention, nonnutritive habits, substance abuse, intraoral/perioral piercing, and speech/language development." (AAPD Clinical Guidelines, p. 116).

Supporting evidence cited by AAPD Guidelines:

American Academy of Pediatrics. Oral health risk assessment timing and establishment of the dental home. Pediatr 2003:11(5):1113-6. Reaffirmed 2009;124(2): 845.

American Academy of Pediatric Dentistry. Guideline on infant oral health care. Pediatr Dent 2012;34 (special issue):132-6.

American Academy of Pediatric Dentistry. Guideline on adolescent oral health care. Pediatr Dent 2012;34(special issue):137-44.

American Academy of Pediatric Dentistry. Policy on prevention of sports-related orofacial injuries. Pediatr Dent 2013;35(special issue):67-71

American Academy of Pediatric Dentistry. Policy on the dental home. Pediatr Dent 2012;34(special issue):24-5.

American Academy of Pediatric Dentistry. Guideline on management of the developing dentition and occlusion in pediatric dentistry. Pediatr Dent 2012;34(special issue):239-51.

CDC. Preventing tobacco use among young people: A report of the Surgeon General (executive summary). MMWR Recommend Reports 1994;43(RR-4):[inclusive page numbers]

American Academy of Pediatric Dentistry. Policy on tobacco use. Pediatr Dent 2012;34(special issue):61-4.

American Academy of Pediatric Dentistry. Policy on intra- oral/perioral piercing and oral jewelry/accessories. Pediatr Dent 2012;34(special issue):65-6.

Douglass JM. Response to Tinanoff and Palmer: Dietary determinants of dental caries and dietary recommendations for preschool children. J Public Health Dent 2000; 60(3):207-9

Kranz S, Smiciklas-Wright H, Francis LA. Diet quality, added sugar, and dietary fiber intakes in American pre-schoolers. Pediatr Dent 2006;28(2):164-71.

Lewis CW, Grossman DC, Domoto PK, Deyo RA. The role of the pediatrician in the oral health of children: A national survey. Pediatrics 2000;106(6):E84.

Li H, Zou Y, Ding G. Dietary factors associated with dental erosion: A meta-analysis. PLoSOne 2012;7(8):e42626. doi:10.1371/journal.pone.0042626. Epub2012 Aug 31.

Malmgren B, Andreasen JO, Flores MT, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Injuries: 3. Injuries in the primary dentition. Dent Traumatol 2012;28(3):174-82. 19.

Mobley C, Marshall TA, Milgrom P, Coldwell SE. The contribution of dietary factors to dental caries and disparities in caries. Acad Pediatr 2009;9(6):410-4

Reisine S, Douglass JM. Pyschosocial and behavorial issues in early childhood caries. Comm Dent Oral Epidem 1998;26(suppl):132-44.

Sigurdsson, A. Evidence-based review of prevention of dental injuries. Pediatr Dent 2013;35(2):184-90.

Tinanoff NT, Palmer C. Dietary determinants of dental caries in pre-school children and dietary recommendations for pre-school children. J Pub Health Dent 2000; 60(3):197-206.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

NICE Guidelines

"The recommended interval between oral health reviews should be determined specifically for each patient and tailored to meet his or her needs, on the basis of an assessment of disease levels and risk of or from dental disease." (NICE Guidelines, 2004, p. 40)

Grade: D

"The shortest interval between oral health reviews for all patients should be 3 months." (NICE Guidelines, 2004, p. 41) Note: NICE uses the term "oral health reviews"

Grade: GPP

"The longest interval between oral health reviews for patients younger than 18 years should be 12 months." (NICE Guidelines, 2004, p. 41)

Grade: GPP

"For practical reasons, the patient should be assigned a recall interval of 3, 6, 9, or 12 months if he or she is younger than 18 years, or 3, 6, 9, 12, 15, 18, 21, or 24 months if he or she is aged 18 years or older." (NICE Guidelines, 2004, p. 41)

Grade: GPP

AAPD Clinical Guidelines

Not graded. Supporting evidence is cited within the guidelines. Please see references in 1a.4.2. above.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

NICE Guidelines (p. 8)

A	> At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population, or > A systematic review of RCTs or a body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results
B	> A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results, or > Extrapolated evidence from studies rated as 1++ or 1+
C	> A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results, or > Extrapolated evidence from studies rated as 2++
D	>Evidence level 3 or 4, or > Extrapolated evidence from studies rated as 2+, or > Formal consensus
GPP	A good practice point (GPP) is a recommendation for best practice based on the clinical experience of the Guideline Development Group

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same as 1a.4.1.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ **X**No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable.

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**. Not applicable.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade: Not applicable.

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*) Not applicable.

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable.

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Riley P, Worthington HV, Clarkson JE, Beirne PV. Recall intervals for oral health in primary care patients. Cochrane Database of Systematic Reviews 2013, Issue 12. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004346.pub4/abstract

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Not applicable.

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

NICE Guidelines

Key Clinical Questions:

(a) How effective are routine dental checks of different recall frequencies in improving quality of life and reducing the morbidity associated with dental caries and periodontal disease in children?

(b) How effective are routine dental checks of different recall frequencies in improving quality of life, reducing the morbidity associated with dental caries, periodontal disease and oral cancer, and reducing the mortality associated with oral cancer in adults?

AAPD Guidelines

The periodicity guideline covers a broad range of services. Consequently, the evidence review for the most recent update of this guideline (2013), included the following search terms for articles published in the last 10 years: “periodicity of dental examinations”, “dental recall intervals”, “preventive dental services”, “anticipatory guidance and dentistry”, “caries risk assessment”, “early childhood caries”, “dental caries prediction”, “dental care cost effectiveness children”, “periodontal disease and children and adolescents US”, “pit and fissure sealants”, “dental sealants“, “fluoride supplementation and topical fluoride”, “dental trauma”, “dental fracture and tooth”, “nonnutritive oral habits”, “treatment of developing malocclusion”, “removal of wisdom teeth”, “removal of third molars”. Additional search limitations were humans, English language, clinical trials, and ages birth -18 years. The search returned 3,418 articles, 113 which were chosen for a detailed review after reviewing the titles and abstracts. (AAPD Clinical Guidelines, p. 114)

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

NICE Guidelines

Although NICE has a detailed method for grading evidence in developing clinical guidelines, the report does not contain the specific grades assigned for the evidence associated with each clinical guideline.

AAPD Guidelines - Evidence grades were not assigned.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

NICE's Evidence Grading System is (p. 6):

1++	High-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
1-	Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias
2++	High-quality systematic reviews of case control or cohort studies High-quality case-control or cohort studies with a very low risk of confounding, bias or chance and a high probability that the relationship is causal
2+	Well-conducted case-control or cohort studies with a low risk of confounding, bias or chance and a moderate probability that the relationship is causal
2-	Case-control or cohort studies with a high risk of confounding bias or chance and a significant risk that the relationship is not causal
3	Non-analytic studies (for example, case reports, case series)
4	Expert opinion, formal consensus

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).

Date range: NICE: NICE built upon an existing systematic review that addressed the focus the guidelines conducted by Davenport et al. (2003). Davenport et al.'s review covered the literature through February 2001. NICE updated that search through July 2003. The AAPD Guidelines conducted a literature search covering the period 2003-2013 for the most recent update of the guidelines; however, evidence from earlier guideline issuance is also included. These guidelines were first adopted in 1991.

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

NICE Guidelines

The literature review addressed a range of outcomes for children and adult associated with different dental recall intervals. There was no restriction on study design. A total of 38 studies were used to make final recommendations. (p.5)

AAPD Guidelines

The AAPD guidelines do not provide a detailed summary of this information. For the update, there were 113 articles selected for detailed review. The search was restricted to clinical trials.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

NICE Guidelines

The guidelines noted a lack of high-quality evidence in this area. However, it also advised: "A recommendation's grade may not necessarily reflect the importance attached to the recommendation. For example, the Guideline Development Group agreed that the principles underlying the individualisation of recall intervals advocated in this guideline are particularly important." (p. 40)

AAPD Guidelines

The guidelines do not provide a formal grade of the quality of evidence across studies. However, these studies were reviewed by dental experts serving on the AAPD's Clinical Affairs Committee and the overall recommendations were further reviewed by the Council on Clinical Affairs. APPD guidelines are developed by members of the AAPD's Council on Clinical Affairs, Council on Scientific Affairs, and additional participants with appropriate expertise. The review team must include members from both academia and clinical practice. Members also participate in evidence-based training sessions sponsored by the AAPD.

Overall Assessment

Although high-quality evidence is lacking, there is expert consensus nationally and internationally based on the best evidence currently available that children should have a routine dental check-up at least once a year and more often based on the individual child's disease and risk status.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Not specifically assessed as part of the review for guideline development. However, as noted above, there is expert consensus regarding the benefits of routine dental check-ups for children at least once per year and more often based on their disease and risk status. In addition, the IOM has identified access to oral health care as a critical first step to improving oral health outcomes and reducing disparities. As demonstrated elsewhere in this application, there are significant performance gaps in use of dental services.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Not specifically assessed as part of the review for guideline development. However, minimal harm would be expected from a routine dental visit.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

A more recent Cochrane review evaluated this topic (Riley et al. 2013). The Cochrane review only included randomized controlled trials; thus, only 1 study was included. The study compared the effects of a clinical examination every 12 months with a clinical examination every 24 months on the outcomes of caries (decayed, missing, filled surfaces (dmfs/DMFS) increment) and economic cost outcomes (total time used per person). The main finding of that study was: "For three to five-year olds with primary teeth, the mean difference (MD) in dmfs increment was -0.90 (95% CI -1.96 to 0.16) in favour of 12-month recall. For 16 to 20-year olds with permanent teeth, the MD in DMFS increment was -0.86 (95% CI -1.75 to 0.03) also in favour of 12-month recall." The quality of the body of evidence was rated as very low because the study was at high risk of bias, had a small sample size and only included low-risk participants. Thus, the review authors concluded: "There is a very low quality body of evidence from one RCT which is insufficient to draw any conclusions regarding the potential beneficial and harmful effects of altering the recall interval between dental check-ups. There is no evidence to support or refute the practice of encouraging patients to attend for dental check-ups at six-monthly intervals." This finding is consistent with those of NICE regarding existing evidence and with the NICE guidelines which advise tailoring recall intervals to individual patient needs within a recommended range of 3 months to 12 months for children. As noted by the NICE and Bright Futures guidelines, although the evidence quality is weak, the need for a comprehensive evaluation of oral health remains critical to improving outcomes.

<u>Citation</u>: Riley P, Worthington HV, Clarkson JE, Beirne PV. Recall intervals for oral health in primary care patients. Cochrane Database of Systematic Reviews 2013, Issue 12.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

Not applicable.

1a.8.2. Provide the citation and summary for each piece of evidence.

Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

4_NQF_Evidence-_utilization.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Inequalities in oral health status and inadequate use of oral health care services are well documented (Dye, Li, and Thorton-Evans 2012; IOM 2011a, 2011b; US DHHS 2010). Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, L i, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009).

Improving access to care through the oral health care delivery system is critical to improving oral health outcomes and addressing oral health disparities. In the IOM report, Improving Access to Oral Health Care for Vulnerable and Underserved Populations, there were four overall conclusions. The first conclusion was: "Improving access to oral health care is a critical and necessary first step to improving oral health outcomes and reducing disparities" (IOM 2011b). Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions. However, there are significant performance gaps and disparities in access. Untreated dental caries occurs among 25% of children living in poverty compared with 10.5% of children living above poverty (Dye, L i, and Thorton-Evans 2012). Approximately 75% of children younger than age 6 years did not have at least one visit to a dentist in the previous year (Edlestein 2009). Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), there are significant variations in use of dental services across states, ranging from approximately 25% to 69% (CMS-416 data, FY 2011). Even among the highest performing states, more than one-fourth of publicly-insured children do not have a dental visit during the year. Similar variation between states is observed among children 0-20 years of age enrolled in commercial dental plans (ADA 2013).

The proposed measure, Utilization of Services – Dental Services, captures whether a child received any dental services during the year and, therefore, also measures access to oral health care – the "critical and necessary first step to improving oral health outcomes and reducing disparities" (IOM 2011b). This measure also includes important stratifications by the children's age.

Utilization of Services allows plans and programs to identify the effectiveness of efforts in improving access to oral health services and target performance improvement initiatives accordingly.

This measure is a program/plan specific measure that contributes to the Healthy People 2020 Objective OH 7 that calls for increasing the proportion of children, adolescents, and adults who used the oral health care system in the past year. This is a leading health indicator.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Below are the testing data and results that met scientific acceptability criteria for endorsement. Because there were no changes in the data source, level of analysis or setting, additional testing has not been conducted.

Data Sources:

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. Data from calendar years 2010 and 2011 were used for all programs except Florida Medicaid. Full-year data for CY 2011 were not available for Florida Medicaid. Therefore, we report only CY 2010 data for Florida Medicaid.

In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011. [Technically, there were three plans represented in the data because Texas CHIP was served by a single dental plan. Since the program=plan in that case, we included it in the "program" level data.]

Below we provide summary data for each of the five programs and two plans individually.

Programs

Our source data for the testing included children 0-20 years in each program. The numbers of children ages 0-20 years enrolled at least one month in each program were as follows :

Texas Medicaid, 2011: 3,544,247
Texas Medicaid, 2010: 3,393,963
Texas CHIP, 2011: 842,454
Texas CHIP, 2010: 786,070
Florida CHIP, 2011: 317,146
Florida CHIP, 2010: 315,975
Commercial, 2011: 184,152
Commercial, 2010: 189,968

Florida Medicaid, 2010: 2,068,670

Within these programs, we had claims data available in both years for two dental managed care plans in Florida CHIP. We also report rates for those two plans separately.

Plan 1, 2010: 77,255
Plan 2, 2010: 116,388
Plan 1, 2011: 140,986
Plan 2, 2011: 168,191

Data 1b.2. Performance Scores for Utilization of Dental Services

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

Program 1, CY 2011:	69.52%	(	0.6952	,	0.0003	,	0.6947	,	0.6957	)
Program 2, CY 2011:	56.34%	(	0.5634	,	0.0007	,	0.5621	,	0.5647	)
Program 3, CY 2011:	52.42%	(	0.5242	,	0.0011	,	0.5221	,	0.5263	)
Program 4, CY 2011:	70.60%	(	0.7060	,	0.0012	,	0.7037	,	0.7083	)
Program 1, CY 2010:	63.13%	(	0.6313	,	0.0003	,	0.6307	,	0.6319	)
Program 2, CY 2010:	54.92%	(	0.5492	,	0.0007	,	0.5478	,	0.5506	)
Program 3, CY 2010:	50.62%	(	0.5062	,	0.0011	,	0.5040	,	0.5084	)
Program 4, CY 2010:	73.81%	(	0.7381	,	0.0012	,	0.7358	,	0.7404	)
Program 5, CY2010:	27.72%	(	0.2772	,	0.0003	,	0.2765	,	0.2779	)
Plan 1, CY 2011:	52.43%	(	0.5243	,	0.0017	,	0.5211	,	0.5275	)
Plan 2, CY 2011:	51.40%	(	0.5140	,	0.0015	,	0.5111	,	0.5169	)
Plan 1, CY 2010:	49.50%	(	0.4950	,	0.0025	,	0.4901	,	0.4999	)
Plan 2, CY 2010 :	47.74%	(	0.4774	,	0.0019	,	0.4737	,	0.4811	)

The measure rate range of 28% to 74% in CY 2010 (year in which data were available for all five programs) indicates a significant performance gap overall. Even in the highest performing program, one-fourth of children did not receive a dental service during the year. In addition, these results demonstrate the ability of the measure to identify variations in performance between programs.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

The measure testing findings are consistent with other data indicating that children have sub-optimal utilization of dental services. Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), there are significant variations in use of dental services across states, ranging from approximately 25% to 69% (CMS EPSDT Data, FY 2011). Even among the highest performing states, more than one-fourth of publicly-insured children do not have a dental visit during the year. Untreated dental caries occurs among 25% of children living in poverty compared with 10.5% of children living above poverty (Dye, L i, and Thorton-Evans 2012). Approximately three quarters of children younger than age 6 years did not have at least one visit to a dentist in the previous year (Edlestein 2009). Similar variation between states is observed among children 0-20 years of age enrolled in commercial dental plans (ADA 2013).

[Complete citations provided in 1c4 and in Evidence Submission Form Template.]

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The data below summarizes performance data by age, geographic location, and race/ethnicity for CY 2011 (CY 2010 for one program) with the p-values from chi-square tests used to detect whether there were statistically significant differences in performance between groups. The results demonstrate that there are disparities by age, geographic location (all except one

program), and race/ethnicity. In addition, we also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), Medicaid program type, commercial product line, and preferred language for program communications. We detected disparities based on each of these various factors, but data on all of these characteristics were not consistently available for all programs so we are presenting disparities data on those characteristics that were most consistently available and had the greatest standardization

Data1b.4. Disparities in Performance by Child Age, Geographic Location and Race/Ethnicity
PROGRAM 1
Overall performance score: 69.52%
Scores by Age
Age <1 years: 18.78%
Age 1-2 years: 59.06%
Age 3-5 years: 75.19%
Age 6-7 years: 78.45%
Age 8-9 years: 78.54%
Age 10-11 years: 77.80%
Age 12-14 years: 77.25%
Age 15-18 years: 69.42%
Age 19-20 years: 42.73%
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 70.48%
Rural: 63.74%
p-value from Chi-square test: <.0001
Scores by Race
Non-Hispanic White: 59.09%
Non-Hispanic Black: 66.09%
Hispanic: 75.18%
p-value from Chi-square test <.0001

PROGRAM 2
Overall performance score: 56.34%
Scores by Age
Age <1 years: 7.17%
Age 1-2 years: 45.95%
Age 3-5 years: 58.27%
Age 6-7 years: 63.55%
Age 8-9 years: 63.49%
Age 10-11 years: 61.17%
Age 12-14 years: 55.56%
Age 15-18 years: 47.38%
Age 19-20 years: N/A
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 57.42%
Rural: 49.79%
p-value from Chi-square test: <.0001
Scores by Race
Non-Hispanic White: N/A
Non-Hispanic Black: N/A
Hispanic: N/A
p-value from Chi-square test N/A

PROGRAM 3
Overall performance score: 52.42%
Scores by Age
Age <1 years: N/A

Age 1-2 years: N/A
Age 3-5 years: 41.28%
Age 6-7 years: 54.17%
Age 8-9 years: 58.40%
Age 10-11 years: 56.09%
Age 12-14 years: 52.58%
Age 15-18 years: 47.43%
Age 19-20 years: N/A
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 52.52%
Rural: 52.11%
p-value from Chi-square test: 0.1393
Scores by Race
Non-Hispanic White: N/A
Non-Hispanic Black: N/A
Hispanic: N/A
p-value from Chi-square test N/A

PROGRAM 4
Overall performance score: 70.60%
Scores by Age
Age <1 years: 0.89%
Age 1-2 years: 13.34%
Age 3-5 years: 66.43%
Age 6-7 years: 80.53%
Age 8-9 years: 82.46%
Age 10-11 years: 80.50%
Age 12-14 years: 78.90%
Age 15-18 years: 72.35%
Age 19-20 years: 62.42%
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 70.93%
Rural: 62.94%
p-value from Chi-square test: <.0001
Scores by Race
Non-Hispanic White: N/A
Non-Hispanic Black: N/A
Hispanic: N/A
p-value from Chi-square test N/A

PROGRAM 5
Overall performance score: 27.72%
Scores by Age
Age <1 years: 0.32%
Age 1-2 years: 6.21%
Age 3-5 years: 28.96%
Age 6-7 years: 38.91%
Age 8-9 years: 41.86%
Age 10-11 years: 38.44%
Age 12-14 years: 34.33%
Age 15-18 years: 29.63%
Age 19-20 years: 18.08%
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 27.09%

Rural: 35.71%
p-value from Chi-square test: <.0001
Scores by Race
Non-Hispanic White: 26.26%
Non-Hispanic Black: 25.52%
Hispanic: 32.18%
p-value from Chi-square test <.0001

Note: N/A for age indicates that those ages are not within the program's age eligibility. N/A for race/ethnicity indicates that those programs did not collect race/ethnicity data or had high rates of missing data.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

There is extensive literature documenting disparities in dental service use among children by age, race/ethnicity, and geographic region, including within vulnerable populations. For example, using data from the National Health and Nutrition Examination Survey, researchers at the National Center for Health Statistics identified variations in untreated dental caries among children by race and ethnicity and poverty level (Dye, L i, and Thorton-Evans 2012). Specifically, they found: "In 2009–2010, 14% of children aged 3–5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries. Among children aged 3–5 years, the prevalence of untreated caries was significantly higher for non-Hispanic black children (19%) compared with non-Hispanic white children (11%). Untreated caries was nearly twice as high for Hispanic children (26%) compared with non-Hispanic white children (14%) aged 6–9 years, and was more than twice as high for non-Hispanic black adolescents (25%) compared with non-Hispanic white adolescents (9%) aged 13–15. For children aged 3–5 and 6–9 years living at or below 100% of the federal poverty level, untreated dental caries was significantly higher compared with children living above the poverty level" (Dye, L i, and Thorton-Evans 2012, pp. 1-2).

Using data from the Medical Expenditure Panel Survey, Edelstein and Chinn (2009, p. 417) noted disparities in dental utilization (any dental visit) by age, family income, race and ethnicity, and education: "Stepwise disparities in dental utilization by income remained as strong in 2004 as in 1996, with 30.8% of poor children, 33.9% of low-income children, 46.5% of middle income children, and 61.8% of high income children having at least 1 dental visit in 2004. One third of minority children (34.1% black and 32.9% of Hispanic children) obtain dental care in a year compared with half (52.5%) of white children. Children whose parents attained less than high school education were less than half as likely to obtain a dental visit in 2004 as children whose parents are college graduates (25% vs 54%)." A recent analysis by Bouchery (2013) of the Medicaid Analytic eXtract files for nine states, examined dental utilization for preventive services and treatment services and found variations in dental service use by age, race, and geographic area. Specifically, relative to the reference group of 9 year olds, the percentage point change in the probability of having a dental preventive services was -27.6 for 3 years old; -8.6 for 6 years, -2.2 for 12 years and -15.4 for 15 years (all significant at $p<0.0001$); relative to the reference group of white, non-Hispanic, the percentage point change was -1.8 for black non-Hispanic and 7.8 for Hispanic ($p<0.0001$ for both); relative to the reference group of small metro area, the percentage point change was 5.9 for large metro area ($p<0.0001$). Relative to the reference group of 9 year olds, the percentage point change in the probability of having a dental treatment services was -19.4 for 3 years old; -8.9 for 6 years, 1.8 for 12 years and -4.3 for 15 years (all significant at $p<0.0001$); relative to the reference group of white, non-Hispanic, the percentage point change was -3.9 for black non-Hispanic and 7.3 for Hispanic ($p<0.0001$ for both); relative to the reference group of small metro area, the percentage point change was 2.9 for large metro area ($p<0.0001$) and -1.2 for noncore adjacent to metro area or micropolitan ($p=0.01$).

Disparities in the use of dental services have also been noted in other literature and summarized in three major national reports on oral health: the Surgeon General's report on Oral Health in America in 2000, the IOM report, Improving Access to Oral Health Care for Vulnerable and Underserved Populations, and the IOM report, Advancing Oral Health in America.

Sources

Blackwell, D. L. 2010. Family structure and children's health in the United States: Findings from the National Health Interview Survey, 2001–2007. Hyattsville, MD: National Center for Health Statistics.

Bouchery, E. 2013. "Utilization of Dental Services among Medicaid-Enrolled Children." Medicare & Medicaid Research Review. 3(3) E1-16. Available at: https://www.cms.gov/mmrr/Downloads/MMRR2013_003_03_b04.pdf.

Dietrich, T., C. Culler, R. Garcia, and M. M. Henshaw. 2008. Racial and ethnic disparities in children's oral health: The National Survey of Children's Health. Journal of the American Dental Association 139(11):1507-1517.

Dye BA, Li X, Thorton-Evans G. Oral health disparities as determined by selected healthy people 2020 oral health objectives for the United States, 2009-2010. NCHS Data Brief 2012(104):1-8.U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research.

Edelstein, B. L. and C. H. Chinn. 2009. "Update on Disparities in Oral Health and Access to Dental Care for America´s Children." Acad Pediatr 9(6): 415-9.

Institute of Medicine (U.S.). Committee on an Oral Health Initiative. Advancing oral health in America. Washington, D.C.: National Academies Press; 2011.

Institute of Medicine and National Research Council. Improving access to oral health care for vulnerable and underserved populations. Washington, D.C.: National Academies Press; 2011.

Manski, R. J., and E. Brown. 2007. Dental use, expenses, private dental coverage, and changes, 1996 and 2004. Rockville, MD: Agency for Healthcare Research and Quality.

U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research. Oral health in America : a report of the Surgeon General. Rockville, Md.: U.S. Public Health Service, Dept. of Health and Human Services; 2000.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Dental

De.6. **Non-Condition Specific***(check all the areas that apply):*
Access to Care, Disparities Sensitive, Health and Functional Status : Change, Health and Functional Status : Total Health, Primary Prevention

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, Populations at Risk

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.ada.org/~/media/ADA/Science%20and%20Research/Files/DQA_2018_Dental_Services_Utilization_of_Services.pdf?la=en

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
1. No changes to the measure specifications

2. Measure specification website updated to be more user friendly

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Unduplicated number of children under age 21 years who received at least one dental service

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see section S.14

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Unduplicated number of enrolled children under age 21 years

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see section S.14

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
There are no other exclusions than those described above.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
This measure is stratified by age using the following categories:

<1; 1-2; 3-5; 6-7; 8-9; 10-11; 12-14; 15-18; 19-20

No new data are needed for this stratification. Please see attached specifications for complete measure details.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Utilization of Services Calculation

1. Use administrative enrollment and claims data for a single year. When using claims data to determine service receipt, include both paid and unpaid claims (including pending, suspended, and denied claims).

2. Check if the enrollee meets age criteria at the last day of the reporting year:
a. If age criterion is met, then proceed to next step.
b. If age criterion is not met or there are missing or invalid field codes (e.g., date of birth), then STOP processing. This enrollee does not get counted in the denominator.

3. Check if subject is continuously enrolled for at least 180 days during the reporting year:
a. If subject meets continuous enrollment criterion, then include in denominator; proceed to next step.
b. If subject does not meet enrollment criterion, then STOP processing. This enrollee does not get counted in the denominator.

YOU NOW HAVE THE DENOMINATOR (DEN) COUNT: All enrollees who meet the age and enrollment criteria

4. Check if subject received any dental service:
a. If [CDT CODE] = D0100 – D9999, and;
b. If [RENDERING PROVIDER TAXONOMY] code = any of the NUCC maintained Provider Taxonomy Codes or their equivalent in Table 1 below, then include in numerator; STOP processing
c. If both a & b are not met, then service was not provided or not a dental service; STOP processing. This enrollee is already included in the denominator but will not be included in the numerators.

Note: In this step, all claims with missing or invalid CDT CODE, missing or invalid NUCC maintained Provider Taxonomy Codes, or NUCC maintained Provider Taxonomy Codes that do not appear in Table 1 should not be included in the numerator.

YOU NOW HAVE NUMERATOR NUM COUNT: Enrollees who received a dental service

5. Report
a. Unduplicated number of enrollees in numerator
b. Unduplicated number of enrollees in denominator
c. Measure Rate (NUM/DEN)
d. Rate stratified by age

Table 1: NUCC maintained Provider Taxonomy Codes classified as "Dental Service"*

122300000X	1223P0106X	1223X0008X	261QF0400X
1223D0001X	1223P0221X	1223X0400X	261QR1300X
1223D0004X	1223P0300X	124Q00000X+	125Q00000X
1223E0200X	1223P0700X	125J00000X	
1223G0001X	1223S0112X	125K00000X	

*Services provided by County Health Department dental clinics may also be included as "dental" services.
+Only dental hygienists who provide services under the supervision of a dentist should be classified as "dental" services. Services provided by independently practicing dental hygienists should be classified as "oral health" services and are not applicable for this measure.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
Not applicable.

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
Not applicable.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
Not applicable.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable.

2. Validity – See attached Measure Testing Submission Form
5_NQF_Testing-utilization.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted

(prior testing as well as any new testing); use red font to indicate updated testing.
No

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Utilization of 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.

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] 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] 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]

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]

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; [14,15] 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] 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
☒ **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:	☐ other:

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
   A. 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
   A. 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
   A. 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
   A. 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
   A. Size: # Members 0-20 years, CY 2010: 2,068,670;
   B. 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

2) FL CHIP – Plan 2
   A. 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 <u>patients</u> 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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2011					
	Program 1	Program 2	Program 3	Program 4	Plan 1	Plan 2
Total Number Patients	3,544,247	842,454	317,146	184,152	140,986	168,191
Age Group Distribution						
Age <1 years	7.05%	0.11%	N/A	1.54%	N/A	N/A
Age 1-2 years	14.32%	5.34%	N/A	5.75%	N/A	N/A
Age 3-5 years	19.46%	11.70%	3.81%	12.68%	4.12%	3.60%
Age 6-7 years	11.21%	12.30%	13.05%	9.57%	13.71%	12.55%
Age 8-9 years	9.85%	14.40%	15.00%	10.18%	15.76%	14.41%
Age 10-11 years	9.03%	14.03%	15.71%	10.55%	16.27%	15.25%
Age 12-14 years	11.63%	19.57%	23.73%	16.09%	23.06%	24.31%
Age 15-18 years	13.19%	22.54%	28.70%	22.13%	27.08%	29.88%
Age 19-20 years	4.27%	N/A	N/A	11.50%	N/A	N/A
Geographic Location						
Urban	83.63%	84.33%	92.94%	95.95%	93.01%	92.91%
Rural	15.15%	14.61%	5.02%	3.86%	4.83%	5.15%
Missing	1.22%	1.06%	2.04%	0.19%	2.16%	1.94%
Race and Ethnicity						
Non-Hispanic White	17.36%	N/A	N/A	N/A	N/A	N/A
Non-Hispanic Black	15.08%	N/A	N/A	N/A	N/A	N/A
Hispanic	58.07%	N/A	N/A	N/A	N/A	N/A
Other & Unknown	9.49%	N/A	N/A	N/A	N/A	N/A

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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2010						
	Program 1	Program 2	Program 3	Program 4	Program 5	Plan 1	Plan 2
Total Number Patients	3,393,963	786,070	315,975	189,968	2,068,670	77,255	116,388
Age Group Distribution							
Age <1 years	7.35%	0.15%	N/A	1.45%	6.05%	N/A	N/A
Age 1-2 years	15.16%	5.37%	N/A	5.67%	14.23%	N/A	N/A
Age 3-5 years	19.48%	11.69%	3.64%	12.73%	19.26%	5.72%	4.22%
Age 6-7 years	11.12%	12.19%	13.32%	9.69%	10.47%	15.68%	12.54%
Age 8-9 years	9.70%	14.61%	15.14%	10.24%	9.19%	16.99%	14.21%
Age 10-11 years	8.75%	14.04%	15.84%	10.60%	8.74%	16.41%	15.18%
Age 12-14 years	11.23%	19.49%	23.70%	16.20%	11.87%	21.40%	24.05%
Age 15-18 years	12.99%	22.47%	28.37%	22.12%	14.73%	23.79%	29.81%
Age 19-20 years	4.22%	N/A	N/A	11.31%	5.47%	N/A	N/A
Geographic Location							
Urban	83.20%	84.46%	92.08%	96.70%	91.47%	92.10%	92.11%
Rural	15.56%	14.45%	5.07%	3.17%	7.30%	5.00%	5.19%
Missing	1.24%	1.08%	2.85%	0.13%	1.23%	2.89%	2.70%
Race and Ethnicity							
Non-Hispanic White	18.21%	N/A	N/A	N/A	29.89%	N/A	N/A
Non-Hispanic Black	15.45%	N/A	N/A	N/A	29.39%	N/A	N/A
Hispanic	59.42%	N/A	N/A	N/A	29.65%	N/A	N/A
Other & Unknown	6.92%	N/A	N/A	N/A	11.06%	N/A	N/A

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) an environmental scan that identified existing measure concepts and their limitations and (2) 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
Utilization of services measures the percentage of children with any dental service using procedure codes in administrative claims data to identify dental services. 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

# Requested	# Received	%Received
600	414	69%

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 specific domains of care representing a range of dental services (e.g., oral evaluation, professionally applied topical fluoride, sealants, and restorations).

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 – Dental Service Procedures Codes

Data Extraction. For the second component of validation, assessing whether specific domains of care performed are accurately captured by claims data, chart abstraction forms were developed by the research team. The specific domains of care evaluated were clinical oral evaluations, topical fluoride, sealants, and restorations. 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 each of the four different types of services (oral evaluations, topical fluoride, sealants, and restorations) 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 corresponding to each of these service categories. 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

	Percentage of All Enrolled Members Included in Different Denominator Definitions				
	Program 1	Program 2	Program 3	Program 4	Program 5
At least 1 month	100%	100%	100%	100%	100%
At least 3 months	95%	85%	84%	93%	94%
At least 6 months	83%	63%	65%	81%	81%
11-12 months	64%	33%	42%	63%	59%

Table 2b2.2-3. Performance Rates for Different Denominator Definitions

	Performance Rates for Different Denominator Definitions				
	Program 1	Program 2	Program 3	Program 4	Program 5
At least 3 months	65%	49%	46%	66%	25%
At least 6 months	70%	56%	52%	71%	28%
11-12 months	76%	65%	57%	75%	31%

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 "any dental services." 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 "any 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

Number of Procedure Codes	Record and Procedure Code on Claim Correlate	Record Did Not Correlate with Procedure Code on Claim	Unable to Determine Correlation
1,135	94.04%	4.22%	1.75%

B. Critical Data Element Validation – Dental Service Procedure Codes for Specific Domains of Care

To assess whether dental services performed are accurately captured by claims data, the 414 records, representing 631 dates of service, were reviewed for all services except topical fluoride. Topical fluoride was not a covered benefit in Texas CHIP during the study period so only Texas Medicaid records, representing 317 dates of service were reviewed. Table 2b2.3-2 below summarizes the agreement between the dental records and administrative data for specific care domains. Agreement ranged from 86.6% to 95.5%, indicating high overall concordance between the administrative claims and dental records. Sensitivity ranged from 77.8% to 90.7%, and specificity ranged from 88.4%-99.3%. Positive predictive values were consistently high, ranging from 93.3% to 98.1%. Negative predictive values ranged from 59.7% for oral evaluation to 95.5% for sealants. 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 care domains for which there was the strongest agreement, based on both simple agreement and the kappa statistic, were sealant applications and restorations. These services had a 95% simple agreement rate and kappa values exceeding 0.80 indicating "almost perfect" agreement. Fluoride applications and oral evaluations both demonstrated "substantial" agreement with overall kappa statistic value of 0.782 and 0.642, respectively, and simple agreement of 89.9% and 86.6%, respectively.

Our findings are similar 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. Comparing billing data to direct observation they found kappa values equal to 0.84 for sealants, 0.81 for fluoride, 0.44 for oral examinations, and 0.79 for amalgam restorations. The main difference between their findings and ours was that they found lower agreement for oral examinations than we did. 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." (p. 472) (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.)

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

	Concordance	Prevalence	Sensitivity	Specificity	PPV	NPV	Kappa
Sealants Applied	95.22%	0.172	0.778	0.988	0.933	0.955	0.820
Dates of service: 613			(0.686-0.850)	(0.974-0.995)	(0.855-0.973)	(0.933-0.971)	(0.758-0.882)
# indeterminate: 4							
Fluoride	89.91%	0.647	0.907	0.884	0.935	0.839	0.782
Dates of service: 317			(0.857-0.942)	(0.806-0.934)	(0.888-0.963)	(0.757-0.898)	(0.710-0.853)
# indeterminate: 0							
Oral Evaluation	86.56%	0.808	0.851	0.925	0.979	0.597	0.6419
Dates of service: 613			(0.817-0.881)	(0.858-0.963)	(0.960-0.990)	(0.522-0.667)	(0.574-0.710)
# indeterminate: 6							
Restorations	95.54%	0.291	0.863	0.993	0.981	0.946	0.888
Dates of service: 613			(0.803-0.908)	(0.979-0.998)	(0.942-0.995)	(0.921-0.964)	(0.848-0.928)
# indeterminate: 3							

95% confidence intervals indicated in parentheses

*Fluoride was not a covered benefit in Texas CHIP, so only Texas Medicaid records were evaluated for this service.

2. FACE VALIDITY

Utilization of Services was identified through the Delphi rating process as a high-scoring measure concept with a mean importance score of 7, mean feasibility score of 8, and mean validity score of 7, 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. These gaps are addressed in more detail in Section 5 (Relation to Other NQF-Endorsed Measures).

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 "any dental services." The rates calculated for the proposed Utilization of Dental Services 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 within 5 percentage points 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 measure, 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 Utilization of Dental Services Score to Similar Domain Calculated using CMS Form 416 EPSDT Data

	Comparison of Measure Score to Similar Domain Calculated using CMS Form 416 EPSDT Data			
	TX Medicaid		FL Medicaid	
	Utilization of Services Measure Score, CY 2011	Percentage of EPSDT Eligibles, CMS-Form 416, FFY 2011	Utilization of Services Measure Score, CY 2010	Percentage of EPSDT Eligibles, CMS-Form 416, FFY 2011
Overall	64.58%	65.45%	25.39%	23.54%
Age Group				
Age <1 years	12.55%	15.28%	0.24%	0.51%
Age 1-2 years	55.71%	56.79%	5.85%	6.98%
Age 3-5 years	71.54%	72.60%	27.29%	27.86%
Age 6-7 years	74.68%	75.75%	36.47%	34.74%
Age 8-9 years	74.69%		39.13%	
Age 10-11 years	73.97%	74.98%	35.79%	29.58%
Age 12-14 years	73.66%		31.83%	
Age 15-18 years	65.26%	66.43%	27.19%	23.81%
Age 19-20 years	38.48%	39.17%	15.68%	12.01%

*Note: DQA age stratifications are more refined than CMS for children in age ranges of 6-9 years and 10-14 years.

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?*)
As noted above, the overall agreement between the administrative claims data and dental record data was high based on both simple agreement and using the more conservative Kappa statistic. We interpret these findings as evidence of strong concurrence between dental records and administrative data. In addition, face validity and convergent validity of the measure scores were established. Collectively, these findings lead us to conclude that both the data elements and the measure score represent valid measures of dental service use.

2b3. EXCLUSIONS ANALYSIS
NA X☐ no exclusions — *skip to section 2b4*
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.
Not applicable.

2b4.1. What method of controlling for differences in case mix is used?
☐X No risk adjustment or stratification
☐ **Statistical risk model with risk factors**
☐ **Stratification by risk categories**
☐ **Other,**

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*) Not applicable.

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

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)

Program 1, CY 2011:	69.52%	(	0.6952 ,	0.0003 ,	0.6947 ,	0.6957)
Program 2, CY 2011:	56.34%	(	0.5634 ,	0.0007 ,	0.5621 ,	0.5647)
Program 3, CY 2011:	52.42%	(	0.5242 ,	0.0011 ,	0.5221 ,	0.5263)
Program 4, CY 2011:	70.60%	(	0.706 ,	0.0012 ,	0.7037 ,	0.7083)
Program 1, CY 2010:	63.13%	(	0.6313 ,	0.0003 ,	0.6307 ,	0.6319)
Program 2, CY 2010:	54.92%	(	0.5492 ,	0.0007 ,	0.5478 ,	0.5506)
Program 3, CY 2010:	50.62%	(	0.5062 ,	0.0011 ,	0.504 ,	0.5084)
Program 4, CY 2010:	73.81%	(	0.7381 ,	0.0012 ,	0.7358 ,	0.7404)
Program 5, CY2010:	27.72%	(	0.2772 ,	0.0003 ,	0.2765 ,	0.2779)
Plan 1, CY 2011:	52.43%	(	0.5243 ,	0.0017 ,	0.5211 ,	0.5275)
Plan 2, CY 2011:	51.40%	(	0.514 ,	0.0015 ,	0.5111 ,	0.5169)
Plan 1, CY 2010:	49.50%	(	0.495 ,	0.0025 ,	0.4901 ,	0.4999)
Plan 2, CY 2010 :	47.74%	(	0.4774 ,	0.0019 ,	0.4737 ,	0.4811)

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

	Chi-Square Value	*p* - value
Program Results, 2011	56427.0	<0.0001
Program Results, 2010	562969.2	<0.0001
Plan Results, 2011	21.1	<0.0001
Plan Results, 2010	32.1	<0.0001

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 regarding use of dental services. 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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
This measure is specified for reporting at program and plan level, and there are currently no plans for developing eMeasures (eCQM) for this measure.

Note for 3b3: Our understanding is that the Feasibility Score Card is only for eMeasures; consequently, we have not submitted this. Feasibility criteria were met during the initial endorsement review.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

This measure relies on standard data elements in administrative claims data (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes. A key advantage of using administrative claims data is that the time and cost of data collection for performance measurement purposes are relatively low because these data are already collected for other purposes.

Initial feasibility assessments were conducted using the RAND-UCLA modified Delphi process to rate the measure concepts with feasibility as one component of the assessment. On a 1-9 point scale, this measure concept was rated as an 8 or "definitely feasible" by the expert panel. During the empirical testing phase, our testing found that the critical data elements had missing/invalid data of <1% (Data 3c.1.), meeting or exceeding the guidance from the Centers for Medicare and Medicaid Services regarding acceptable error rates. During measure development and testing, the measure specifications were made available through a publicly accessible website for public comment with additional broad email dissemination to a wide range of stakeholders. No concerns regarding feasibility were raised during this process.

Citation: Centers for Medicare & Medicaid Services. Medicaid and CHIP Statistical Information System (MSIS) File Specifications and Data Dictionary. 2010; http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MSIS/downloads/msisdd2010.pdf. Accessed August 10, 2013.

Data 3c.1 Percentage of Missing and Invalid Values for Critical Data Elements

PROGRAM 1
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.01%
Rendering Provider ID: 0.28%

PROGRAM 2
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.00%
Rendering Provider ID: 0.00%

PROGRAM 3
Member ID: 0.27%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.28%
Date of Service: 0.00%
Rendering Provider ID: 0.18%

PROGRAM 4
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.01%
Date of Service: 0.00%
Rendering Provider ID: 0.61%

PROGRAM 5
Member ID: 0.43%
Date of Birth: 0.02%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.00%
Rendering Provider ID: 0.67%

Endorsement Maintenance Update: There have been no reports of feasibility issues with implementing this measure. Please see Use and Usability section.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
This measure is intended to be transparent and available for widespread adoption. As such, it was purposefully designed to avoid using software or other proprietary materials that would require licensing fees. The measure specifications, including a companion User Guide, are accessible through a website and can be used free of charge for non-commercial purposes. The main requirement of users is to ensure the quality of their source data and expertise to program the measures within their information systems, following the clear and detailed specifications. Technical assistance is available to users.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting Texas Medicaid and CHIP https://hhs.texas.gov/sites/default/files/documents/laws-regulations/handbooks/umcm/10-1-10.pdf Quality Improvement (external benchmarking to organizations) Covered California http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf Michigan Healthy Kids Dental RFP

	https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids Texas Medicaid and CHIP https://hhs.texas.gov/sites/default/files/documents/laws-regulations/handbooks/umcm/10-1-10.pdf Quality Improvement (Internal to the specific organization) State Medicaid Agencies http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

4a1.1 **For each CURRENT use, checked above (update for <u>maintenance of endorsement</u>), provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

1. Covered California, the California Health Benefit Exchange

http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf
http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-QDP-Issuer-Contract-and-Attachments.pdf

Purpose: Quality Improvement

This measure is included in the Covered California Qualified Health Plan Issuer Contract for 2017-019 For the Individual Market and the Covered California Qualified Dental Plan Issuer Contract for 2017-2019. The measure is to be reported annually.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies statewide. In March 2017 there were 85,000 enrollees 0-18 years old in CC health plans (which may offer dental benefits and would therefore report on the dental quality measures). There were 5,100 children enrolled specifically in Qualified Dental Plans. (http://hbex.coveredca.com/data-research/)

Level of Measurement and Setting. The measure is implemented at the plan level with the Covered California program.

2. State Medicaid Agencies

http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

(Note: To access the data, a public user account must be created. We can help facilitate access to the data if needed.)

Purpose: Quality Improvement

The Medicaid|Medicare|CHIP Services Dental Association conducts an annual survey of state Medicaid programs and collects data specifically on which programs report Dental Quality Alliance measures.

In its 2015 profile (the most recent available), 15 states reported that they currently use this measure in the Medicaid and/or CHIP programs.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The 15 states are: Alabama, California, Colorado, Connecticut, Florida, Idaho, Illinois, Louisiana, Michigan, Nevada, Oklahoma, Rhode Island, South Carolina, Virginia and West Virginia. Data are not provided on the number of accountable entities included.

3. Michigan Healthy Kids Dental Program

https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids

Note: Select Schedule A Work Statement link under File Attachments

Purpose: Quality Improvement

The Michigan Healthy Kids Dental Program has included this measure in the set of measures included in its Performance Monitoring Standards, which is currently included in the Request for Proposals and will be included in the contracts between the contracted dental plans and the State of Michigan.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The Healthy Kids Dental Program covers children enrolled in Michigan's Medicaid program statewide. The state intends to award two contracts. There are approximately 955,000 enrollees served by the Healthy Kids Dental Program.

4. Texas Health and Human Services Commission – Texas Medicaid and CHIP

https://hhs.texas.gov/sites/default/files/documents/laws-regulations/handbooks/umcm/10-1-10.pdf
and
https://hhs.texas.gov/sites/default/files/documents/laws-regulations/handbooks/umcm/10-1-9.pdf

Purpose: Quality Improvement and Public Reporting

This measure has been adopted by the Texas Health and Human Services Commission as part of the Texas CHIP and Medicaid Dental Services Performance Indicator Dashboard for Quality Measures Program. [Texas HHSC Uniform Managed Care Manual, Chapters 10.1.9 and 10.1.10. Effective Date 01/15/2016, Version 2.5].

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies to the state of Texas CHIP and Medicaid programs (statewide application). There are two dental plans (i.e., the accountable entities) that serve Texas CHIP and Medicaid. In June 2017, there were 3,359,770 children enrolled in Texas Medicaid and CHIP (https://hhs.texas.gov/about-hhs/records-statistics/data-statistics/healthcare-statistics).

Level of Measurement and Setting: The measure is implemented at the plan and program level within the Texas Medicaid and CHIP programs.

Additional Information:
This measure was one of ten performance measures that focused on Dental Caries Prevention and Disease Management among children approved by the DQA. 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)

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not applicable.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Per the annual survey conducted by the Medicaid|Medicare|CHIP Services Dental Association (MSDA), 15 Medicaid/CHIP agencies are implementing this measure. The measure is part of measure set included in the Request for Proposal (RFP) released by the Michigan Healthy Kids Dental Program. This measure is included in the Texas Medicaid/CHIP performance dashboard. Additionally, this measure is a requirement for the Qualified Dental Plans to report to the Covered California, the state-based marketplace in California.

The DQA provides technical assistance to these and other users of DQA measures through webinars, resource document development, and one-on-one staff support. The DQA has an Implementation Committee dedicated to developing implementation and improvement resources.

In order to ensure transparency, incorporate learnings from implementation, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting. Implementation has mostly focused on addressing questions related to how to use the measures in the context of broader quality improvement and clarifying questions related to the specifications.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**

In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

The DQA provides technical assistance to users of DQA measures on an ongoing basis through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**

Describe how feedback was obtained.

In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

The DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.3. **Summarize the feedback obtained from other users**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures either have only limited baseline scores or will start reporting measures within the next year.

We are only aware of repeat measurements within the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after it was approved by the Dental Quality Alliance and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 2698361, 69.61, 71.02, 68.28
2015, 2929975, 71.49, 72.70, 69.97

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 452976, 61.96, 64.62, 61.67
2015, 341937, 65.90, 70.44, 67.36

These data suggest a trend in improvement over time. However, as noted above, these are initial performance data for one program. Most measure users are just now getting their quality measurement programs underway.

4b2. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
There are no unexpected findings.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization of Related Measures
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific

submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Appendix_UtilServices.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Dental Association on behalf of the Dental Quality Alliance
Co.2 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-
Co.3 **Measure Developer if different from Measure Steward:** American Dental Association on behalf of the Dental Quality Alliance
Co.4 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This project is headed by the DQA through its Measure Development and Maintenance Committee (formerly Research and Development Committee). The following individuals were responsible for executing and overseeing all scientific aspects of this project.

- Craig W. Amundson, DDS, General Dentist, HealthPartners, National Association of Dental Plans. Dr. Amundson serves as chair for the Committee.
- Mark Casey, DDS, MPH, Dental Director, North Carolina Department of Health and Human Services Division of Medical Assistance
- Natalia Chalmers, DDS, PhD, Diplomate, American Board of Pediatric Dentistry, Director, Analytics and Publication, DentaQuest Institute
- Frederick Eichmiller, DDS, Vice President & Science Officer, Delta Dental of Wisconsin
- Chris Farrell, RDH, BSDH, MPA, Oral Health Program Director, Michigan Department of Health and Human Services

This group oversees the maintenance of the measures. All work of this Committee was distributed for review and formal vote and approval by the entire Dental Quality Alliance. (http://ada.org/dqa) The DQA is made up of representatives from 38 stakeholder organizations.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2013
Ad.3 **Month and Year of most recent revision:** 01, 2017
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2018

Ad.7 **Disclaimers:** Dental Quality Alliance measures and related data specifications, developed by the Dental Quality Alliance (DQA), are intended to facilitate quality improvement activities. These Measures are intended to assist stakeholders in enhancing quality of care. These performance Measures are not clinical guidelines and do not establish a standard of care. The DQA has not tested its Measures for all potential applications.
Measures are subject to review and may be revised or rescinded at any time by the DQA. The Measures may not be altered without the prior written approval of the DQA. The DQA shall be acknowledged as the measure steward in any and all references to the measure.
Measures developed by the DQA, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and DQA. Neither the DQA nor its members shall be responsible for any use of these Measures.
THE MEASURES ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND
Limited proprietary coding is contained in the Measure specifications for convenience.
For Proprietary Codes:

Ad.8 **Additional Information/Comments:** In 2008, the Centers for Medicare and Medicaid Services (CMS) asked the ADA to lead the development of a broad coalition of organizations that would lead dentistry to improve the oral health of Americans through quality measurement and quality improvement. The ADA subsequently established the DQA. The DQA is a multi-stakeholder alliance comprised of approximately 38 stakeholders (with organizations as members) from across the oral health community, including federal agencies, third-party payers, professional associations, and an individual member from the general public. The DQA's mission is to advance the field of performance measurement to improve oral health, patient care, and safety through a consensus building process.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2517
Measure Title: Oral Evaluation, Dental Services
Measure Steward: American Dental Association on behalf of the Dental Quality Alliance
Brief Description of Measure: Percentage of enrolled children under age 21 years who received a comprehensive or periodic oral evaluation within the reporting year.
Developer Rationale: Inequalities in oral health status and inadequate use of oral health care services are well documented (Dye, Li, and Thorton-Evans 2012; IOM 2011a, 2011b; US DHHS 2010). Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, Li, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009).

Identifying dental caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions. Comprehensive and periodic clinical oral evaluations are diagnostic services that are critical to evaluating oral disease and dentition development.* Clinical oral evaluations also are essential to developing an appropriate preventive oral health regimen and treatment plan. Thus, clinical oral evaluations play an essential role in caries identification, prevention and treatment, thereby promoting improved oral health, overall health, and quality of life.

National guidelines from the American Academy of Pediatric Dentistry (AAPD) and the American Academy of Pediatrics (AAP) recommend that children receive oral health services by 1 year of age and have regular visits thereafter. The most common recall interval is six months. However, evidence-based guidelines indicate that the recall schedule for routine oral evaluations should be tailored to individual needs based on assessments of existing disease and risk of disease (e.g., caries risk) with a recommended recall frequency ranging from 3 months to no more than 12 months for individuals younger than 18 years of age (National Institute for Health and Care Excellence (NICE), Clinical Guideline 19, 2004).

However, there are significant performance gaps and disparities in care. Untreated dental caries occurs among 25% of children living in poverty compared with 10.5% of children living above poverty (Dye, Li, and Thorton-Evans 2012). Approximately 75% of children younger than age 6 years did not have at least one visit to a dentist in the previous year (Edelstein and Chinn 2009) despite the recommendation that every child have a visit by 12 months of age. Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), 23% to 63% of children enrolled in Medicaid/CHIP for at least 90 continuous days receive an oral evaluation (referred to as "Dental Diagnostic Services") (CMS EPSDT Data, FY 2011). Even among the highest performing states, more than one-third of publicly-insured children do not receive an oral evaluation as a dental service during the year. Thus, a significant percentage of children are not receiving oral evaluations to assess their oral health status and disease risk and develop an appropriate preventive oral health regimen and treatment plan tailored to individual needs.

The proposed measure, Oral Evaluation - Dental Services, captures whether children receive a comprehensive or periodic oral evaluation as a dental service during the reporting year. In addition, this measure also includes important stratifications by the children's age. Oral Evaluation allows plans and programs to assess whether children are receiving at least one oral evaluation during the reporting year as recommended by evidence-based guidelines.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

* A Comprehensive Oral Evaluation may be performed on new or established patients and is "a thorough evaluation and recording of the extraoral and intraoral hard and soft tissues" and includes "an evaluation for oral cancer where indicated, the evaluation and recording of the patient's dental and medical history and a general health assessment. It may include the evaluation and recording of dental caries, missing or unerupted teeth, restorations, existing prostheses, occlusal relationships, periodontal conditions (including periodontal screening and/or charting), hard and soft tissue anomalies, etc." A Periodic Oral Evaluation is performed "on a patient of record to determine any changes in the patient's dental and medical health status since a previous comprehensive or periodic evaluation." In addition, there is a code for Oral Evaluation for a Patient under Three Years of Age and Counseling with Primary Caregiver, which includes "[d]iagnostic services performed for a child under the age of three, preferably within the first six months of the eruption of the first primary tooth, including recording of the oral and physical health history, evaluation of caries susceptibility, development of an appropriate preventive oral health regimen and communication with and counseling of the child's parent, legal guardian and/or primary caregiver." American Dental Association. 2012. "CDT 2013: Dental Procedure Codes." Chicago, IL: American Dental Association.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

Numerator Statement: Unduplicated number of enrolled children under age 21 years who received a comprehensive or periodic oral evaluation as a dental service

Denominator Statement: Unduplicated number of enrolled children under age 21 years

Denominator Exclusions: Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded

Measure Type: Process

Data Source: Claims

Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Sep 18, 2014 **Most Recent Endorsement Date:** Sep 18, 2014

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- Clinical oral evaluations play an essential role in caries identification, prevention and treatment, thereby promoting improved oral health, overall health, and quality of life. Evidence-based guidelines recommend clinical oral evaluations with a regular recall schedule that is tailored to individual needs based on assessments of existing disease and risk of disease (e.g., caries risk) with the recommended recall frequency ranging from 3 months to no more than 12 months for individuals younger than 18 years of age (National Institute for Health and Care Excellence (NICE), Clinical Guideline 19, 2004).
- NICE Guidelines: Although NICE has a detailed method for grading evidence in developing clinical guidelines, the report does not contain the specific grades assigned for the evidence associated with each clinical guideline.
- AAPD Guidelines: Evidence grades were not assigned.

Citations:

National Institute for Health and Care Excellence (NICE). 2004. Clinical Guidelines. "CG19: Dental Recall – Recall Interval between Routine Dental Examinations." Available at: http://guidance.nice.org.uk/CG19.

American Academy of Pediatric Dentistry. 2013. "Guideline on Periodicity of Examination, Preventive Dental Services, Anticipatory Guidance/Counseling, and Oral Treatment for Infants, Children, and Adolescents. " Available at: http://www.aapd.org/media/Policies_Guidelines/G_Periodicity.pdf.

American Academy of Pediatrics Section on Pediatric Dentistry and Oral Health. 2008. "Policy Statement: Preventive Oral Health Intervention for Pediatricians." Pediatrics 122(6): 1387-94. Available at: http://pediatrics.aappublications.org/content/122/6/1387.full.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- A more recent Cochrane review evaluated this topic (Riley et al. 2013). The Cochrane review only included randomized controlled trials; thus, only one study was included. The main finding of that study was: "For three to five-year olds with primary teeth, the mean difference (MD) in dmfs increment was -0.90 (95% CI -1.96 to 0.16) in favour of 12-month recall. For 16 to 20-year olds with permanent teeth, the MD in DMFS increment was -0.86 (95% CI -1.75 to 0.03) also in favour of 12-month recall."

Citation:

Riley P, Worthington HV, Clarkson JE, Beirne PV. Recall intervals for oral health in primary care patients. Cochrane Database of Systematic Reviews 2013, Issue 12.

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review but does note a recent Cochrane review collated all evidence and reached the same conclusions that supported the original guideline. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm
Process measure based on systematic review (Box 3)→ Empirical evidence submitted (Box 7)→ Empirical evidence includes all studies in body of evidence (Box 8) → Rate as Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer used data from five sources and refers to "program" level information and "plan" level information (Texas Medicaid, Texas CHIP, Florida CHIP, and Florida Medicaid programs as well as national commercial data from Dental Service of Massachusetts, Inc.). The developer presented the total number of children enrolled in each program/plan. In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011.
- The data source and sample size are sufficient to assess gaps in performance. The performance range of **26% to 67% in CY 2010** (year in which data were available for all four programs) indicates a significant performance gap overall. With respect to oral evaluations specifically, 23% to 63% of children enrolled in Medicaid/CHIP for at least 90 continuous days receive an oral evaluation (referred to as "Dental Diagnostic Services") (CMS EPSDT Data, FY 2011).
- The developer did not provide more recent performance data, stating that due to the start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting.

Disparities

- The developer found disparities based by age, geographic location, and race/ethnicity. In addition, it also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), CHIP dental plan, Medicaid program type, commercial product line, and preferred language for program communications. The developer detected disparities based on each of these various factors, but data on all of these characteristics were not consistently available for all programs so we are presenting disparities data on those characteristics that were most consistently available and had the greatest standardization (i.e. race/ethnicity and geographic location).

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.
2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Staff Scientific Acceptability Rating Logic

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure relies on **standard data elements in administrative claims data** (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes.

- ***Update:*** The developer states there have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on

performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- Texas Health and Human Services Commission: Medicaid/CHIP Pay For Quality Program (P4Q) https://hhs.texas.gov/sites/default/files//documents/lawsregulations/ handbooks/umcm/6-2-15.pdf

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- In 2016, the Dental Quality Alliance (DQA) expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from six state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from eightdental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

The developer notes that it is only aware of repeat measurements within the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after it was approved by the Dental Quality Alliance and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 2698361, 67.35, 69.23, 65.39
2015, 2929975, 69.12, 71.21, 66.49

Texas CHIP

Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 452976, 59.43, 62.90, 58.23
2015, 341937, 63.41, 68.79, 63.62

The developer notes that these data suggest a trend in improvement over time. However, as noted above, these are initial performance data for one program. Most measure users are just now getting their quality measurement programs underway.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation
No unintended or negative consequences were identified by the developer.

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- N/A

Harmonization

- N/A

Committee pre-evaluation comments
Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Staff Scientific Acceptability Rating Logic

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?
   ☒Yes (go to Question #2)
   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?
   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)
   ☒Yes (go to Question #4)
   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?
   ☒Yes (use your rating from data element validity testing – Question #16- under Validity Section)
   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?
   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
   ☐Yes (go to Question #5)
   ☒No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☐Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☐High (go to Question #8)

☐Moderate (go to Question #8)
☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☒Yes (go to Question #9)
   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☒Yes (go to Question #10)
   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?
   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.
   ☒Yes (go to Question #2)
   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?
   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?
   ☐Yes (please explain below then go to Question #3)
   ☐No (go to Question #3)
   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?
   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*
   ☐Yes (please explain below then go to Question #4)
   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
   ☐Yes (please explain below then go to Question #5)
   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
   ☐Yes (please explain below then go to Question #6)
   ☒No (go to Question #6)

☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☐Yes (go to Question #11)
    ☒No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☐Yes (go to Question #12)

☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
    ☐High (go to Question #14)
    ☐Moderate (go to Question #14)
    ☐Low (please explain below then go to Question #13)
    ☐Insufficient

13. Was other validity testing reported?
    ☒Yes (go to Question #14)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?
    TIPS: Prior validity studies of the same data elements may be submitted
    ☒Yes (go to Question #15)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*
    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
    ☒Yes (go to Question #16)
    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
    ☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
    ☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Oral Evaluation, Dental Services

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 2/10/2014

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☐ Health outcome:

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome:

X☐ Process: Receipt of a comprehensive or periodic oral evaluation during the reporting period

☐ Structure:

☐ Other:

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not applicable.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The proposed measure, Oral Evaluation - Dental Services, captures whether children receive a comprehensive or periodic oral evaluation as a dental service during the reporting year. As described in 1b1 (Importance), dental caries is the most common chronic disease in children in the U.S., and a significant percentage of children have untreated dental caries. Dental decay causes significant short- and long-term adverse consequences for children's health and functioning. Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions. Evidence-based guidelines

recommend clinical oral evaluations with a regular recall schedule that is tailored to individual needs based on assessments of existing disease and risk of disease (e.g., caries risk) with the recommended recall frequency ranging from 3 months to no more than 12 months for individuals younger than 18 years of age (National Institute for Health and Care Excellence (NICE), Clinical Guideline 19, 2004). Comprehensive and periodic clinical oral evaluations are diagnostic services that are critical to evaluating oral disease and dentition development. Clinical oral evaluations also are essential to developing an appropriate preventive oral health regimen and treatment plan. Thus, clinical oral evaluations play an essential role in caries identification, prevention and treatment, thereby promoting improved oral health, overall health, and quality of life.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ **Clinical Practice Guideline recommendation** – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

National Institute for Health and Care Excellence (NICE). 2004. Clinical Guidelines. "CG19: Dental Recall – Recall Interval between Routine Dental Examinations." Available at: http://guidance.nice.org.uk/CG19.

American Academy of Pediatric Dentistry. 2013. "Guideline on Periodicity of Examination, Preventive Dental Services, Anticipatory Guidance/Counseling, and Oral Treatment for Infants, Children, and Adolescents. " Available at: http://www.aapd.org/media/Policies_Guidelines/G_Periodicity.pdf.

American Academy of Pediatrics Section on Pediatric Dentistry and Oral Health. 2008. "Policy Statement: Preventive Oral Health Intervention for Pediatricians." Pediatrics 122(6): 1387-94. Available at: http://pediatrics.aappublications.org/content/122/6/1387.full.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

National guidelines from the American Academy of Pediatric Dentistry (AAPD) and the American Academy of Pediatrics (AAP) recommend that children receive oral health services by 1 year of age and have regular visits thereafter. The most common recall interval is six months. However, evidence-based guidelines indicate that the recall schedule should be tailored to individual needs based on assessments of existing disease and risk of disease (e.g., caries risk) with a recommended recall frequency for routine oral evaluations ranging from 3 months to no more than 12 months for individuals younger than 18 years of age.

Terminology Note: The United Kingdom's National Institute for Health and Care Excellence (NICE) uses the term "Oral Health Review" to "refer to the continuing re-examination of an individual's oral health and risk status." The UK's Oral Health Reviews are what the American Dental Association refers to as "Oral Evaluations."

Age of First Visit

"The first examination is recommended at the time of the eruption of the first tooth and no later than 12 months of age." (p. 114 of AAPD Clinical Guidelines).

"Every child should have a dental home established by 1 year of age." (American Academy of Pediatrics Section on Pediatric Dentistry and Oral Health. 2008. "Policy Statement: Preventive Oral Health Intervention for Pediatricians." Pediatrics 122(6): 1387-94; at page 1391).

Supporting evidence cited in AAPD Guidelines:

American Academy of Pediatric Dentistry. Policy on the dental home. Pediatr Dent 2012;34(special issue):24-5.

American Academy of Pediatrics. Oral health risk assessment timing and establishment of the dental home. Pediatr 2003:11(5):1113-6. Reaffirmed 2009;124(2):

Berg JH, Stapleton FB. Physician and dentist: New initiatives to jointly mitigate early childhood oral disease. Clin Pediatr 2012:51(6):531-7.

Recall Interval

"The recommended interval between oral health reviews should be determined specifically for each patient and tailored to meet his or her needs, on the basis of an assessment of disease levels and risk of or from dental disease." (NICE Guidelines, 2004, p. 40)

"The shortest interval between oral health reviews for all patients should be 3 months." (NICE Guidelines, 2004, p. 41) Note: NICE uses the term "oral health reviews"

"The longest interval between oral health reviews for patients younger than 18 years should be 12 months." (NICE Guidelines, 2004, p. 41)

- Rationale: "There is evidence that the rate of progression of dental caries can be more rapid in children and adolescents than in older people, and it seems to be faster in primary teeth than in permanent teeth (see Chapter Three, Section 3.1.2.) Periodic developmental assessment of the dentition is also required in children. Recall intervals of no longer than 12 months give the opportunity for delivering and reinforcing preventive advice and for raising awareness of the importance of good oral health. This is particularly important in young children, to layout the foundations for life-long dental health." (NICE Guidelines, 2004, p. 41)

"For practical reasons, the patient should be assigned a recall interval of 3, 6, 9, or 12 months if he or she is younger than 18 years, or 3, 6, 9, 12, 15, 18, 21, or 24 months if he or she is aged 18 years or older." (NICE Guidelines, 2004, p. 41)

"The most common interval of examination is six months; however, some patients may require examination and preventive services at more or less frequent intervals, based upon historical, clinical, and radiographic findings." (p. 115 of AAPD Clinical Guidelines)

Supporting evidence cited by AAPD Clinical Guidelines:

Beil HA, Rozier RG. Primary health care providers' advice for a dental checkup and dental use in children. Pediatr 2010;126(2):435-41.

Pahel BT, Rozier RG, Stearns SC, Quiñonez RB. Effectiveness of preventive dental treatments by physicians for young Medicaid enrollees. Pediatr 2011;127(3):682-9.

Diangelis AJ, Andreasen JO, Ebeleseder KA, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Dental Injuries: 1. Fractures and luxations of permanent teeth. Dent Traumatol 2012;28(1):2-12.

Andersson L, Andreasen JO, Day P, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Dental Injuries: 2. Avulsion of permanent teeth. Dent Traumatol 2012;28(2):88-96.

Malmgren B, Andreasen JO, Flores MT, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Injuries: 3. Injuries in the primary dentition. Dent Traumatol 2012;28(3):174-82.

Patel S, Bay RC, Glick M. A systematic review of dental recall intervals and incidence of dental caries. J Am Dent Assoc 2010;141(5):527-39.

American Academy of Pediatric Dentistry. Guideline on prescribing dental radiographs. Pediatr Dent 2012;34(special issue):299-301.

American Dental Association Council on Scientific Affairs.The use of dental radiographs; Update and recommendations. J Am Dent Assoc 2006;137(9):1304-12.

Greenwell H, Committee on Research, Science and Therapy American Academy of Periodontology. Guidelines for periodontal therapy. J Periodontol 2001;72(11):1624-8.

Califano JV, Research Science and Therapy CommitteeAmerican Academy of Periodontology. Periodontal diseases of children and adolescents. J Periodontol 2003;74(11):1696-704.

Clerehugh V. Periodontal diseases in children and adoles¬cents. British Dental J 2008;204(8):469-71.845.

Benefits Obtained

"Early detection and management of oral conditions can improve a child's oral health, general health and well-being, and school readiness." (p. 114 of AAPD Clinical Guidelines)

Supporting evidence cited by AAPD Guidelines:

American Academy of Pediatric Dentistry. Policy on early childhood caries: Classifications, consequences, and preventive strategies. Pediatr Dent 2012;34(special issue):50-2.

American Academy of Pediatric Dentistry. Policy on early childhood caries: Unique challenges and treatment options. Pediatr Dent 2012;34(special issue):53-5.

Clarke M, Locker D, Berall G, Pencharz P, Kenny DJ, Judd P. Malnourishment in a population of young children with severe early childhood caries. Pediatr Dent 2006;28(3):254-9.

Dye BA, Shenkin JD, Ogden CL, Marshall TA, Levy SM, Kanellis MJ. The relationship between healthful eating practices and dental caries in children ages 2-5 years in the United States, 1988-1994. J Am Dent Assoc 2004;135(1):55-6.

Jackson SL, Vann WF, Kotch J, Pahel BT, Lee JY. Impact of poor oral health on children's school attendance and performance. Amer J Publ Health 2011;10(10):1900-6.

Every visit provides the opportunity to provide anticipatory guidance, which "is the process of providing practice, developmentally-appropriate information about children's health to prepare parents for the significant physical, emotional, and psychological milestones." (AAPD Clinical Guidelines, p. 116) "Individualized discussion and counseling [anticipatory guidance] should be an integral part of each visit. Topics to be included are oral hygiene and dietary habits, injury prevention, nonnutritive habits, substance abuse, intraoral/perioral piercing, and speech/language development." (AAPD Clinical Guidelines, p. 116).

Supporting evidence cited by AAPD Guidelines:

American Academy of Pediatrics. Oral health risk assessment timing and establishment of the dental home. Pediatr 2003:11(5):1113-6. Reaffirmed 2009;124(2): 845.

American Academy of Pediatric Dentistry. Guideline on infant oral health care. Pediatr Dent 2012;34 (special issue):132-6.

American Academy of Pediatric Dentistry. Guideline on adolescent oral health care. Pediatr Dent 2012;34(special issue):137-44.

American Academy of Pediatric Dentistry. Policy on prevention of sports-related orofacial injuries. Pediatr Dent 2013;35(special issue):67-71

American Academy of Pediatric Dentistry. Policy on the dental home. Pediatr Dent 2012;34(special issue):24-5.

American Academy of Pediatric Dentistry. Guideline on management of the developing dentition and occlusion in pediatric dentistry. Pediatr Dent 2012;34(special issue):239-51.

CDC. Preventing tobacco use among young people: A report of the Surgeon General (executive summary). MMWR Recommend Reports 1994;43(RR-4):[inclusive page numbers]

American Academy of Pediatric Dentistry. Policy on tobacco use. Pediatr Dent 2012;34(special issue):61-4.

American Academy of Pediatric Dentistry. Policy on intra- oral/perioral piercing and oral jewelry/accessories. Pediatr Dent 2012;34(special issue):65-6.

Douglass JM. Response to Tinanoff and Palmer: Dietary determinants of dental caries and dietary recommendations for preschool children. J Public Health Dent 2000; 60(3):207-9

Kranz S, Smiciklas-Wright H, Francis LA. Diet quality, added sugar, and dietary fiber intakes in American pre-schoolers. Pediatr Dent 2006;28(2):164-71.

Lewis CW, Grossman DC, Domoto PK, Deyo RA. The role of the pediatrician in the oral health of children: A national survey. Pediatrics 2000;106(6):E84.

Li H, Zou Y, Ding G. Dietary factors associated with dental erosion: A meta-analysis. PLoSOne 2012;7(8):e42626. doi:10.1371/journal.pone.0042626. Epub2012 Aug 31.

Malmgren B, Andreasen JO, Flores MT, et al. International Association of Dental Traumatology Guidelines for the Management of Traumatic Injuries: 3. Injuries in the primary dentition. Dent Traumatol 2012;28(3):174-82. 19.

Mobley C, Marshall TA, Milgrom P, Coldwell SE. The contribution of dietary factors to dental caries and disparities in caries. Acad Pediatr 2009;9(6):410-4

Reisine S, Douglass JM. Pyschosocial and behavorial issues in early childhood caries. Comm Dent Oral Epidem 1998;26(suppl):132-44.

Sigurdsson, A. Evidence-based review of prevention of dental injuries. Pediatr Dent 2013;35(2):184-90.

Tinanoff NT, Palmer C. Dietary determinants of dental caries in pre-school children and dietary recommendations for pre-school children. J Pub Health Dent 2000; 60(3):197-206.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

<u>NICE Guidelines</u>

"The recommended interval between oral health reviews should be determined specifically for each patient and tailored to meet his or her needs, on the basis of an assessment of disease levels and risk of or from dental disease." (NICE Guidelines, 2004, p. 40)

Grade: D

"The shortest interval between oral health reviews for all patients should be 3 months." (NICE Guidelines, 2004, p. 41) Note: NICE uses the term "oral health reviews"

Grade: GPP

"The longest interval between oral health reviews for patients younger than 18 years should be 12 months." (NICE Guidelines, 2004, p. 41)

Grade: GPP

"For practical reasons, the patient should be assigned a recall interval of 3, 6, 9, or 12 months if he or she is younger than 18 years, or 3, 6, 9, 12, 15, 18, 21, or 24 months if he or she is aged 18 years or older." (NICE Guidelines, 2004, p. 41)

Grade: GPP

AAPD Clinical Guidelines

Not graded. Supporting evidence is cited within the guidelines. Please see references in 1a.4.2. above.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

NICE Guidelines (p. 8)

A	> At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population, or > A systematic review of RCTs or a body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results
B	> A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results, or > Extrapolated evidence from studies rated as 1++ or 1+
C	> A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results, or > Extrapolated evidence from studies rated as 2++
D	>Evidence level 3 or 4, or > Extrapolated evidence from studies rated as 2+, or > Formal consensus
GPP	A good practice point (GPP) is a recommendation for best practice based on the clinical experience of the Guideline Development Group

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same as 1a.4.1.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable.

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**. Not applicable.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade: Not applicable.

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Not applicable.

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable.

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Riley P, Worthington HV, Clarkson JE, Beirne PV. Recall intervals for oral health in primary care patients. Cochrane Database of Systematic Reviews 2013, Issue 12. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004346.pub4/abstract

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Not applicable.

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

NICE Guidelines

Key Clinical Questions:

(a) How effective are routine dental checks of different recall frequencies in improving quality of life and reducing the morbidity associated with dental caries and periodontal disease in children?

(b) How effective are routine dental checks of different recall frequencies in improving quality of life, reducing the morbidity associated with dental caries, periodontal disease and oral cancer, and reducing the mortality associated with oral cancer in adults?

AAPD Guidelines

The periodicity guideline covers a broad range of services. Consequently, the evidence review for the most recent update of this guideline (2013), included the following search terms for articles published in the last 10

years: “periodicity of dental examinations”, “dental recall intervals”, “preventive dental services”, “anticipatory guidance and dentistry”, “caries risk assessment”, “early childhood caries”, “dental caries prediction”, “dental care cost effectiveness children”, “periodontal disease and children and adolescents US”, “pit and fissure sealants”, “dental sealants“, “fluoride supplementation and topical fluoride”, “dental trauma”, “dental fracture and tooth”, “nonnutritive oral habits”, “treatment of developing malocclusion”, “removal of wisdom teeth”, “removal of third molars”. Additional search limitations were humans, English language, clinical trials, and ages birth -18 years. The search returned 3,418 articles, 113 which were chosen for a detailed review after reviewing the titles and abstracts. (AAPD Clinical Guidelines, p. 114)

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

NICE Guidelines

Although NICE has a detailed method for grading evidence in developing clinical guidelines, the report does not contain the specific grades assigned for the evidence associated with each clinical guideline.

AAPD Guidelines

Evidence grades were not assigned.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

NICE’s Evidence Grading System is (p. 6):

1++	High-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
1-	Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias
2++	High-quality systematic reviews of case control or cohort studies High-quality case-control or cohort studies with a very low risk of confounding, bias or chance and a high probability that the relationship is causal
2+	Well-conducted case-control or cohort studies with a low risk of confounding, bias or chance and a moderate probability that the relationship is causal
2-	Case-control or cohort studies with a high risk of confounding bias or chance and a significant risk that the relationship is not causal
3	Non-analytic studies (for example, case reports, case series)
4	Expert opinion, formal consensus

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).

Date range: NICE: NICE built upon an existing systematic review that addressed the focus the guidelines conducted by Davenport et al. (2003). Davenport et al.’s review covered the literature through February 2001. NICE updated that search through July 2003. The AAPD Guidelines conducted a literature search covering the period 2003-2013 for the most recent update of the guidelines; however, evidence from earlier guideline issuance is also included. These guidelines were first adopted in 1991.

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

NICE Guidelines

The literature review addressed a range of outcomes for children and adult associated with different dental recall intervals. There was no restriction on study design. A total of 38 studies were used to make final recommendations. (p.5)

AAPD Guidelines

The AAPD guidelines do not provide a detailed summary of this information. For the update, there were 113 articles selected for detailed review. The search was restricted to clinical trials.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

NICE Guidelines

The guidelines noted a lack of high-quality evidence in this area. However, it also advised: "A recommendation's grade may not necessarily reflect the importance attached to the recommendation. For example, the Guideline Development Group agreed that the principles underlying the individualisation of recall intervals advocated in this guideline are particularly important." (p. 40)

AAPD Guidelines

The guidelines do not provide a formal grade of the quality of evidence across studies. However, these studies were reviewed by dental experts serving on the AAPD's Clinical Affairs Committee and the overall recommendations were further reviewed by the Council on Clinical Affairs. APPD guidelines are developed by members of the AAPD's Council on Clinical Affairs, Council on Scientific Affairs, and additional participants with appropriate expertise. The review team must include members from both academia and clinical practice. Members also participate in evidence-based training sessions sponsored by the AAPD.

Overall Assessment

Although high-quality evidence is lacking, there is expert consensus nationally and internationally based on the best evidence currently available that children should have a routine dental check-up (i.e., Oral Evaluation) at least once a year and more often based on the individual child's disease and risk status.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Not specifically assessed as part of the review for guideline development. However, as noted above, there is expert consensus regarding the benefits of routine dental check-ups – Oral Evaluation – for children at least once per year and more often based on their disease and risk status.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Not specifically assessed as part of the review for guideline development. However, minimal harm would be expected from an oral evaluation that involves visual inspection of the oral tissues, evaluation/recording of medical and oral health history, and evaluation for caries risk and risk assessment.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

A more recent Cochrane review evaluated this topic (Riley et al. 2013). The Cochrane review only included randomized controlled trials; thus, only 1 study was included. The study compared the effects of a clinical examination every 12 months with a clinical examination every 24 months on the outcomes of caries (decayed, missing, filled surfaces (dmfs/DMFS) increment) and economic cost outcomes (total time used per person). The main finding of that study was: "For three to five-year olds with primary teeth, the mean difference (MD) in dmfs increment was -0.90 (95% CI -1.96 to 0.16) in favour of 12-month recall. For 16 to 20-year olds with permanent teeth, the MD in DMFS increment was -0.86 (95% CI -1.75 to 0.03) also in favour of 12-month recall." The quality of the body of evidence was rated as very low because the study was at high risk of bias, had a small sample size and only included low-risk participants. Thus, the review authors concluded: "There is a very low quality body of evidence from one RCT which is insufficient to draw any conclusions regarding the potential beneficial and harmful effects of altering the recall interval between dental check-ups. There is no evidence to support or refute the practice of encouraging patients to attend for dental check-ups at six-monthly intervals." This finding is consistent with those of NICE regarding existing evidence and with the NICE guidelines which advise tailoring recall intervals to individual patient needs within a recommended range of 3 months to 12 months for children. As noted by the NICE and Bright Futures guidelines, although the quality of evidence is weak, the need for a comprehensive evaluation of oral health remains critical to improving outcomes. Citation: Riley P, Worthington HV, Clarkson JE, Beirne PV. Recall intervals for oral health in primary care patients. Cochrane Database of Systematic Reviews 2013, Issue 12.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

Not applicable.

1a.8.2. Provide the citation and summary for each piece of evidence.

Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

4_NQF_Evidence-_oral_eval.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Inequalities in oral health status and inadequate use of oral health care services are well documented (Dye, Li, and Thorton-Evans 2012; IOM 2011a, 2011b; US DHHS 2010). Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, Li, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009).

Identifying dental caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions. Comprehensive and periodic clinical oral evaluations are diagnostic services that are critical to evaluating oral disease and dentition development.* Clinical oral evaluations also are essential to developing an appropriate preventive oral health regimen and treatment plan. Thus, clinical oral evaluations play an essential role in caries identification, prevention and treatment, thereby promoting improved oral health, overall health, and quality of life.

National guidelines from the American Academy of Pediatric Dentistry (AAPD) and the American Academy of Pediatrics (AAP) recommend that children receive oral health services by 1 year of age and have regular visits thereafter. The most common recall interval is six months. However, evidence-based guidelines indicate that the recall schedule for routine oral evaluations should be tailored to individual needs based on assessments of existing disease and risk of disease (e.g., caries risk) with a recommended recall frequency ranging from 3 months to no more than 12 months for individuals younger than 18 years of age (National Institute for Health and Care Excellence (NICE), Clinical Guideline 19, 2004).

However, there are significant performance gaps and disparities in care. Untreated dental caries occurs among 25% of children living in poverty compared with 10.5% of children living above poverty (Dye, Li, and Thorton-Evans 2012). Approximately 75% of children younger than age 6 years did not have at least one visit to a dentist in the previous year (Edelstein and Chinn 2009) despite the recommendation that every child have a visit by 12 months of age. Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), 23% to 63% of children enrolled in Medicaid/CHIP for at least 90 continuous days receive an oral evaluation (referred to as "Dental Diagnostic Services") (CMS EPSDT Data, FY 2011).

Even among the highest performing states, more than one-third of publicly-insured children do not receive an oral evaluation as a dental service during the year. Thus, a significant percentage of children are not receiving oral evaluations to assess their oral health status and disease risk and develop an appropriate preventive oral health regimen and treatment plan tailored to individual needs.

The proposed measure, Oral Evaluation - Dental Services, captures whether children receive a comprehensive or periodic oral evaluation as a dental service during the reporting year. In addition, this measure also includes important stratifications by the children's age. Oral Evaluation allows plans and programs to assess whether children are receiving at least one oral evaluation during the reporting year as recommended by evidence-based guidelines.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

* A Comprehensive Oral Evaluation may be performed on new or established patients and is "a thorough evaluation and recording of the extraoral and intraoral hard and soft tissues" and includes "an evaluation for oral cancer where indicated, the evaluation and recording of the patient's dental and medical history and a general health assessment. It may include the evaluation and recording of dental caries, missing or unerupted teeth, restorations, existing prostheses, occlusal relationships, periodontal conditions (including periodontal screening and/or charting), hard and soft tissue anomalies, etc." A Periodic Oral Evaluation is performed "on a patient of record to determine any changes in the patient's dental and medical health status since a previous comprehensive or periodic evaluation." In addition, there is a code for Oral Evaluation for a Patient under Three Years of Age and Counseling with Primary Caregiver, which includes "[d]iagnostic services performed for a child under the age of three, preferably within the first six months of the eruption of the first primary tooth, including recording of the oral and physical health history, evaluation of caries susceptibility, development of an appropriate preventive oral health regimen and communication with and counseling of the child's parent, legal guardian and/or primary caregiver." American Dental Association. 2012. "CDT 2013: Dental Procedure Codes." Chicago, IL: American Dental Association.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

1b.2. Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis. *(<u>This is required for maintenance of endorsement</u>. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Below are the testing data and results that met scientific acceptability criteria for endorsement. Because there were no changes in the data source, level of analysis or setting, additional testing has not been conducted.

Data Sources:

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. Data from calendar years 2010 and 2011 were used for all programs except Florida Medicaid. Full-year data for CY 2011 were not available for Florida Medicaid. Therefore, we report only CY 2010 data for Florida Medicaid.

In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Texas CHIP, (3) Florida CHIP, (4) Commercial Data, and (5) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011. [Technically, there were three plans represented in the data because Texas CHIP was served by a single dental plan. Since the program=plan in that case, we included it in the "program" level data.]

Below we provide summary data for each of the five programs and two plans individually.

Programs

Our source data for the testing included children 0-20 years in each program. The numbers of children ages 0-20 years enrolled at least one month in each program were as follows:

Texas Medicaid, 2011: 3,544,247
Texas Medicaid, 2010: 3,393,963
Texas CHIP, 2011: 842,454
Texas CHIP, 2010: 786,070
Florida CHIP, 2011: 317,146
Florida CHIP, 2010: 315,975
Commercial, 2011: 184,152
Commercial, 2010: 189,968
Florida Medicaid, 2010: 2,068,670

Within these programs, we had claims data available in both years for two dental managed care plans in Florida CHIP. We also report rates for those two plans separately.

Plan 1, 2010: 77,255
Plan 2, 2010: 116,388
Plan 1, 2011: 140,986
Plan 2, 2011: 168,191

Data 1b.2. Performance Scores for Oral Evaluation, Dental Services

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

Program 1, CY 2011:	66.55%	(	0.6655	,	0.0003	,	0.6650	,	0.6660	)
Program 2, CY 2011:	54.18%	(	0.5418	,	0.0007	,	0.5405	,	0.5431	)
Program 3, CY 2011:	46.43%	(	0.4643	,	0.0011	,	0.4622	,	0.4664	)
Program 4, CY 2011:	63.26%	(	0.6326	,	0.0012	,	0.6302	,	0.6350	)
Program 1, CY 2010:	60.59%	(	0.6059	,	0.0003	,	0.6053	,	0.6065	)
Program 2, CY 2010:	52.48%	(	0.5248	,	0.0007	,	0.5234	,	0.5262	)
Program 3, CY 2010:	44.91%	(	0.4491	,	0.0011	,	0.4470	,	0.4512	)
Program 4, CY 2010:	66.96%	(	0.6696	,	0.0012	,	0.6672	,	0.6720	)
Program 5, CY2010:	26.25%	(	0.2625	,	0.0003	,	0.2618	,	0.2632	)
Plan 1, CY 2011:	46.37%	(	0.4637	,	0.0017	,	0.4605	,	0.4669	)
Plan 2, CY 2011:	45.44%	(	0.4544	,	0.0015	,	0.4515	,	0.4573	)
Plan 1, CY 2010:	43.72%	(	0.4372	,	0.0025	,	0.4324	,	0.4420	)
Plan 2, CY 2010 :	41.68%	(	0.4168	,	0.0019	,	0.4132	,	0.4204	)

The measure rate range of 26% to 67% in CY 2010 (year in which data were available for all four programs) indicates a significant performance gap overall. Even in the highest performing program, one-third of children did not receive a comprehensive or period oral evaluation during the year. In addition, these results demonstrate the ability of the measure to identify variations in performance between programs.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

The measure testing findings are consistent with other data indicating that children have sub-optimal utilization of dental services in general and oral evaluations in particular. Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), there are significant variations in use of dental services overall across states, ranging from approximately 25% to 69% (CMS EPSDT Data, FY 2011). Similar variation between states is observed among children 0-20 years of age enrolled in commercial dental plans (ADA 2013). With respect to oral evaluations specifically, 23% to 63% of children

enrolled in Medicaid/CHIP for at least 90 continuous days receive an oral evaluation (referred to as "Dental Diagnostic Services") (CMS EPSDT Data, FY 2011). Even among the highest performing states, more than one-third of publicly-insured children do not receive an oral evaluation as a dental service during the year.

[Complete citations provided in 1c4 and in Evidence Submission Form Template.]

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The same data sources were used as described in 1b.2. The data below summarizes performance data by age, geographic location, and race/ethnicity for CY 2011 (CY 2010 for one program) with the p-values from chi-square tests used to detect whether there were statistically significant differences in performance between groups. The results demonstrate that there are disparities by age, geographic location, and race/ethnicity. In addition, we also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), CHIP dental plan, Medicaid program type, commercial product line, and preferred language for program communications. We detected disparities based on each of these various factors, but data on all of these characteristics were not consistently available for all programs so we are presenting disparities data on those characteristics that were most consistently available and had the greatest standardization.

Data1b.4. Disparities in Performance by Child Age, Geographic Location and Race/Ethnicity

PROGRAM 1

Overall performance score: 66.55%

Scores by Age

- Age <1 years: 18.66%
- Age 1-2 years: 58.83%
- Age 3-5 years: 73.56%
- Age 6-7 years: 76.26%
- Age 8-9 years: 76.24%
- Age 10-11 years: 75.12%
- Age 12-14 years: 71.46%
- Age 15-18 years: 61.99%
- Age 19-20 years: 36.71%

p-value from Chi-square test: <.0001

Scores by Geographic Location

Urban: 67.60%

Rural: 60.10%

p-value from Chi-square test: <.0001

Scores by Race

Non-Hispanic White: 55.80%

Non-Hispanic Black: 62.72%

Hispanic: 72.32%

p-value from Chi-square test <.0001

PROGRAM 2

Overall performance score: 54.18%

Scores by Age

- Age <1 years: 7.17%
- Age 1-2 years: 45.38%
- Age 3-5 years: 56.93%
- Age 6-7 years: 61.33%
- Age 8-9 years: 60.98%
- Age 10-11 years: 59.03%
- Age 12-14 years: 53.37%
- Age 15-18 years: 44.80%

Age 19-20 years: n/a
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 55.40%
Rural: 46.75%
p-value from Chi-square test: <.0001
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 3 46.43%
Overall performance score:
Scores by Age
Age <1 years: n/a
Age 1-2 years: n/a
Age 3-5 years: 39.34%
Age 6-7 years: 50.37%
Age 8-9 years: 53.29%
Age 10-11 years: 50.66%
Age 12-14 years: 46.29%
Age 15-18 years: 39.79%
Age 19-20 years: n/a
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 46.56%
Rural: 45.39%
p-value from Chi-square test: 0.0191
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 4
Overall performance score: 63.26%
Scores by Age
Age <1 years: 0.80%
Age 1-2 years: 11.88%
Age 3-5 years: 62.25%
Age 6-7 years: 75.01%
Age 8-9 years: 75.53%
Age 10-11 years: 73.50%
Age 12-14 years: 70.16%
Age 15-18 years: 63.11%
Age 19-20 years: 52.32%
p-value from Chi-square test: <.0001
Scores by Geographic Location
Urban: 63.61%
Rural: 55.29%
p-value from Chi-square test: <.0001
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 5

Overall performance score:	26.25%
Scores by Age	
Age <1 years:	0.27%
Age 1-2 years:	5.84%
Age 3-5 years:	27.99%
Age 6-7 years:	37.32%
Age 8-9 years:	40.10%
Age 10-11 years:	36.69%
Age 12-14 years:	32.31%
Age 15-18 years:	27.06%
Age 19-20 years:	15.73%
p-value from Chi-square test:	<.0001
Scores by Geographic Location	
Urban:	25.56%
Rural:	34.89%
p-value from Chi-square test:	<.0001
Scores by Race	
Non-Hispanic White:	25.00%
Non-Hispanic Black:	24.18%
Hispanic:	30.35%
p-value from Chi-square test	<.0001

Note: N/A for age indicates that those ages are not within the program's age eligibility. N/A for race/ethnicity indicates that those programs did not collect race/ethnicity data or had high rates of missing data.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

There is extensive literature documenting disparities in dental service use among children by age, race/ethnicity, and geographic region, including within vulnerable populations. For example, using data from the National Health and Nutrition Examination Survey, researchers at the National Center for Health Statistics identified variations in untreated dental caries among children by race and ethnicity and poverty level (Dye, Li, and Thorton-Evans 2012). Specifically, they found: "In 2009–2010, 14% of children aged 3–5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries. Among children aged 3–5 years, the prevalence of untreated caries was significantly higher for non-Hispanic black children (19%) compared with non-Hispanic white children (11%). Untreated caries was nearly twice as high for Hispanic children (26%) compared with non-Hispanic white children (14%) aged 6–9 years, and was more than twice as high for non-Hispanic black adolescents (25%) compared with non-Hispanic white adolescents (9%) aged 13–15. For children aged 3–5 and 6–9 years living at or below 100% of the federal poverty level, untreated dental caries was significantly higher compared with children living above the poverty level" (Dye, Li, and Thorton-Evans 2012, pp. 1-2).

Using data from the Medical Expenditure Panel Survey, Edelstein and Chinn (2009, p. 417) noted disparities in dental utilization (any dental visit) by age, family income, race and ethnicity, and education: "Stepwise disparities in dental utilization by income remained as strong in 2004 as in 1996, with 30.8% of poor children, 33.9% of low-income children, 46.5% of middle income children, and 61.8% of high income children having at least 1 dental visit in 2004. One third of minority children (34.1% black and 32.9% of Hispanic children) obtain dental care in a year compared with half (52.5%) of white children. Children whose parents attained less than high school education were less than half as likely to obtain a dental visit in 2004 as children whose parents are college graduates (25% vs 54%)." A recent analysis by Bouchery (2013) of the Medicaid Analytic eXtract files for nine states, examined dental utilization for preventive services and found variations in dental service use by age, race, and geographic area. Specifically, relative to the reference group of 9 year olds, the percentage point change in the probability of having a dental preventive services was -27.6 for 3 years old; -8.6 for 6 years, -2.2 for 12 years and -15.4 for 15 years (all significant at $p<0.0001$); relative to the reference group of white, non-Hispanic, the percentage point change was -1.8 for black non-Hispanic and 7.8 for Hispanic ($p<0.0001$ for both); relative to the reference group of small metro area, the percentage point change was 5.9 for large metro area ($p<0.0001$). Disparities in the use of dental services have also been noted in other literature and summarized in three major national reports on oral health: the Surgeon General's report on Oral Health in America in 2000, the IOM report, Improving Access to Oral Health Care for Vulnerable and Underserved Populations, and the IOM report, Advancing Oral Health in America.

Sources
Blackwell, D. L. 2010. Family structure and children's health in the United States: Findings from the National Health Interview Survey, 2001–2007. Hyattsville, MD: National Center for Health Statistics.

Bouchery, E. 2013. "Utilization of Dental Services among Medicaid-Enrolled Children." Medicare & Medicaid Research Review. 3(3) E1-16. Available at: https://www.cms.gov/mmrr/Downloads/MMRR2013_003_03_b04.pdf.

Dietrich, T., C. Culler, R. Garcia, and M. M. Henshaw. 2008. Racial and ethnic disparities in children's oral health: The National Survey of Children's Health. Journal of the American Dental Association 139(11):1507-1517.

Dye BA, Li X, Thorton-Evans G. Oral health disparities as determined by selected healthy people 2020 oral health objectives for the United States, 2009-2010. NCHS Data Brief 2012(104):1-8.U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research.

Edelstein, B. L. and C. H. Chinn. 2009. "Update on Disparities in Oral Health and Access to Dental Care for America´s Children." Acad Pediatr 9(6): 415-9.

Institute of Medicine (U.S.). Committee on an Oral Health Initiative. Advancing oral health in America. Washington, D.C.: National Academies Press; 2011.

Institute of Medicine and National Research Council. Improving access to oral health care for vulnerable and underserved populations. Washington, D.C.: National Academies Press; 2011.

Manski, R. J., and E. Brown. 2007. Dental use, expenses, private dental coverage, and changes, 1996 and 2004. Rockville, MD: Agency for Healthcare Research and Quality.

U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research. Oral health in America : a report of the Surgeon General. Rockville, Md.: U.S. Public Health Service, Dept. of Health and Human Services; 2000.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Dental

De.6. **Non-Condition Specific***(check all the areas that apply):*
Access to Care, Disparities Sensitive, Health and Functional Status : Change, Health and Functional Status : Total Health, Primary Prevention, Screening

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, Populations at Risk

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.ada.org/~/media/ADA/Science%20and%20Research/Files/DQA_2018_Oral_Evaluation.pdf?la=en

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
1. No changes to the measure specifications
2. Measure specification website updated to be more user friendly

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Unduplicated number of enrolled children under age 21 years who received a comprehensive or periodic oral evaluation as a dental service

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see Section S14.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Unduplicated number of enrolled children under age 21 years

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see Section S14.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
There are no other exclusions than those described above.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
This measure will be stratified by age using the following categories:

<1; 1-2; 3-5; 6-7; 8-9; 10-11; 12-14; 15-18; 19-20

No new data are needed for this stratification. Please see attached specifications for complete measure details.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Oral Evaluation Calculation

1. Use administrative enrollment and claims data for a single year. When using claims data to determine service receipt, include both paid and unpaid claims (including pending, suspended, and denied claims).

2. Check if the enrollee meets age criteria at the last day of the reporting year:
a. If age criterion is met, then proceed to next step.
b. If age criterion is not met or there are missing or invalid field codes (e.g., date of birth), then STOP processing. This enrollee does not get counted in the denominator.

3. Check if subject is continuously enrolled for at least 180 days:
a. If subject meets continuous enrollment criterion, then include in denominator; proceed to next step.
b. If subject does not meet enrollment criterion, then STOP processing. This enrollee does not get counted in the denominator.

YOU NOW HAVE THE DENOMINATOR (DEN) COUNT: All enrollees who meet age and enrollment criteria

4. Check if subject received an oral evaluation as a dental service:
a. If [CDT CODE] = D0120 or D0150 or D0145, and;
b. If [RENDERING PROVIDER TAXONOMY] code = any of the NUCC maintained Provider Taxonomy Codes in Table 1 below, then include in numerator; proceed to next step.
c. If both a AND b are not met, then the service was not provided or not a "dental service"; STOP processing. This enrollee is already included in the denominator but will not be included in the numerator.

Note: In this step, all claims with missing or invalid CDT CODE, missing or invalid NUCC maintained Provider Taxonomy Codes, or NUCC maintained Provider Taxonomy Codes that do not appear in Table 1 should not be included in the numerator.

YOU NOW HAVE NUMERATOR (NUM) COUNT: Enrollees who received an oral evaluation as a dental service

5. Report
a. Unduplicated number of enrollees in numerator
b. Unduplicated number of enrollees in denominator
c. Measure Rate NUM/DEN
d. Rate stratified by age

Table 1: NUCC maintained Provider Taxonomy Codes classified as "Dental Service"*

122300000X	1223P0106X	1223X0008X	261QF0400X
1223D0001X	1223P0221X	1223X0400X	261QR1300X
1223D0004X	1223P0300X	124Q00000X+	125Q00000X
1223E0200X	1223P0700X	125J00000X	
1223G0001X	1223S0112X	125K00000X	

*Services provided by County Health Department dental clinics may also be included as "dental" services.
+Only dental hygienists who provide services under the supervision of a dentist should be classified as "dental" services. Services provided by independently practicing dental hygienists should be classified as "oral health" services and are not applicable for this measure.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
Not applicable.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
Not applicable.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
Not applicable.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

S.22. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable.

2. Validity – See attached Measure Testing Submission Form
5_Testing-_oral_eval.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of

the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

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.

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] 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] 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]

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]

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; [14,15] 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] **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
☒ 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:	☐ other:

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
   A. 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
   A. 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
   A. 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
   A. 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
   A. Size: # Members 0-20 years, CY 2010: 2,068,670;
   B. 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

2) FL CHIP – Plan 2
   A. 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 <u>patients</u> 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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2011					
	Program 1	Program 2	Program 3	Program 4	Plan 1	Plan 2
Total Number Patients	3,544,247	842,454	317,146	184,152	140,986	168,191
Age Group Distribution						
Age <1 years	7.05%	0.11%	N/A	1.54%	N/A	N/A
Age 1-2 years	14.32%	5.34%	N/A	5.75%	N/A	N/A
Age 3-5 years	19.46%	11.70%	3.81%	12.68%	4.12%	3.60%
Age 6-7 years	11.21%	12.30%	13.05%	9.57%	13.71%	12.55%
Age 8-9 years	9.85%	14.40%	15.00%	10.18%	15.76%	14.41%
Age 10-11 years	9.03%	14.03%	15.71%	10.55%	16.27%	15.25%
Age 12-14 years	11.63%	19.57%	23.73%	16.09%	23.06%	24.31%
Age 15-18 years	13.19%	22.54%	28.70%	22.13%	27.08%	29.88%
Age 19-20 years	4.27%	N/A	N/A	11.50%	N/A	N/A
Geographic Location						
Urban	83.63%	84.33%	92.94%	95.95%	93.01%	92.91%
Rural	15.15%	14.61%	5.02%	3.86%	4.83%	5.15%
Missing	1.22%	1.06%	2.04%	0.19%	2.16%	1.94%
Race and Ethnicity						
Non-Hispanic White	17.36%	N/A	N/A	N/A	N/A	N/A
Non-Hispanic Black	15.08%	N/A	N/A	N/A	N/A	N/A
Hispanic	58.07%	N/A	N/A	N/A	N/A	N/A
Other & Unknown	9.49%	N/A	N/A	N/A	N/A	N/A

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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2010						
	Program 1	Program 2	Program 3	Program 4	Program 5	Plan 1	Plan 2
Total Number Patients	3,393,963	786,070	315,975	189,968	2,068,670	77,255	116,388
Age Group Distribution							
Age <1 years	7.35%	0.15%	N/A	1.45%	6.05%	N/A	N/A
Age 1-2 years	15.16%	5.37%	N/A	5.67%	14.23%	N/A	N/A
Age 3-5 years	19.48%	11.69%	3.64%	12.73%	19.26%	5.72%	4.22%
Age 6-7 years	11.12%	12.19%	13.32%	9.69%	10.47%	15.68%	12.54%
Age 8-9 years	9.70%	14.61%	15.14%	10.24%	9.19%	16.99%	14.21%
Age 10-11 years	8.75%	14.04%	15.84%	10.60%	8.74%	16.41%	15.18%
Age 12-14 years	11.23%	19.49%	23.70%	16.20%	11.87%	21.40%	24.05%
Age 15-18 years	12.99%	22.47%	28.37%	22.12%	14.73%	23.79%	29.81%
Age 19-20 years	4.22%	N/A	N/A	11.31%	5.47%	N/A	N/A
Geographic Location							
Urban	83.20%	84.46%	92.08%	96.70%	91.47%	92.10%	92.11%
Rural	15.56%	14.45%	5.07%	3.17%	7.30%	5.00%	5.19%
Missing	1.24%	1.08%	2.85%	0.13%	1.23%	2.89%	2.70%
Race and Ethnicity							
Non-Hispanic White	18.21%	N/A	N/A	N/A	29.89%	N/A	N/A
Non-Hispanic Black	15.45%	N/A	N/A	N/A	29.39%	N/A	N/A
Hispanic	59.42%	N/A	N/A	N/A	29.65%	N/A	N/A
Other & Unknown	6.92%	N/A	N/A	N/A	11.06%	N/A	N/A

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

# Requested	# Received	%Received
600	414	69%

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

	Percentage of All Enrolled Members Included in Different Denominator Definitions				
	Program 1	Program 2	Program 3	Program 4	Program 5
At least 1 month	100%	100%	100%	100%	100%
At least 3 months	95%	85%	84%	93%	94%
At least 6 months	83%	63%	65%	81%	81%
11-12 months	64%	33%	42%	63%	59%

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

	Performance Rates for Different Denominator Definitions				
	Program 1	Program 2	Program 3	Program 4	Program 5
At least 3 months	62%	47%	41%	59%	24%
At least 6 months	67%	54%	46%	63%	26%
11-12 months	73%	62%	51%	67%	29%

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

Number of Procedure Codes	Record and Procedure Code on Claim Correlate	Record Did Not Correlate with Procedure Code on Claim	Unable to Determine Correlation
1,135	94.04%	4.22%	1.75%

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

	Concordance	Prevalence	Sensitivity	Specificity	PPV	NPV	Kappa
Oral Evaluation	86.56%	0.808	0.851	0.925	0.979	0.597	0.6419
Dates of service: 613			(0.817-0.881)	(0.858-0.963)	(0.960-0.990)	(0.522-0.667)	(0.574-0.710)
# indeterminate: 6							

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

	Comparison of Measure Score to Similar Domain Calculated using CMS Form 416 EPSDT Data			
	TX Medicaid		FL Medicaid	
	Oral Evaluation, Dental Measure Score, CY 2011	Percentage of EPSDT Eligibles, CMS-Form 416, FFY 2011	Oral Evaluation, Dental Measure Score, CY 2010	Percentage of EPSDT Eligibles, CMS-Form 416, FFY 2010
Overall	61.47%	63.05%	24.00%	22.51%
Age Group				
Age <1 years	12.44%	15.24%	0.20%	0.49%
Age 1-2 years	55.47%	56.65%	5.50%	6.76%
Age 3-5 years	69.87%	71.15%	26.35%	26.85%
Age 6-7 years	72.33%	73.57%	34.93%	33.39%
Age 8-9 years	72.24%		37.42%	
Age 10-11 years	71.13%	70.44%	34.12%	28.23%
Age 12-14 years	67.52%		29.92%	
Age 15-18 years	57.60%	59.63%	24.79%	22.37%
Age 19-20 years	32.10%	34.02%	13.56%	11.14%
*Note: DQA age stratifications are more refined than CMS for children in age ranges of 6-9 years and 10-14 years.				

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*
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.
Not applicable.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with risk factors**
☐ **Stratification by risk categories**
☐ **Other,**

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

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)

Program 1, CY 2011:	66.55%	(	0.6655 ,	0.0003 ,	0.6650 ,	0.6660)
Program 2, CY 2011:	54.18%	(	0.5418 ,	0.0007 ,	0.5405 ,	0.5431)
Program 3, CY 2011:	46.43%	(	0.4643 ,	0.0011 ,	0.4622 ,	0.4664)
Program 4, CY 2011:	63.26%	(	0.6326 ,	0.0012 ,	0.6302 ,	0.6350)
Program 1, CY 2010:	60.59%	(	0.6059 ,	0.0003 ,	0.6053 ,	0.6065)
Program 2, CY 2010:	52.48%	(	0.5248 ,	0.0007 ,	0.5234 ,	0.5262)
Program 3, CY 2010:	44.91%	(	0.4491 ,	0.0011 ,	0.4470 ,	0.4512)
Program 4, CY 2010:	66.96%	(	0.6696 ,	0.0012 ,	0.6672 ,	0.6720)
Program 5, CY2010:	26.25%	(	0.2625 ,	0.0003 ,	0.2618 ,	0.2632)
Plan 1, CY 2011:	46.37%	(	0.4637 ,	0.0017 ,	0.4605 ,	0.4669)
Plan 2, CY 2011:	45.44%	(	0.4544 ,	0.0015 ,	0.4515 ,	0.4573)
Plan 1, CY 2010:	43.72%	(	0.4372 ,	0.0025 ,	0.4324 ,	0.4420)
Plan 2, CY 2010 :	41.68%	(	0.4168 ,	0.0019 ,	0.4132 ,	0.4204)

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

	Chi-Square Value	*p* - value
Program Results, 2011	57891.00	<0.0001
Program Results, 2010	521345.50	<0.0001
Plan Results, 2011	17.32	<0.0001
Plan Results, 2010	43.89	<0.0001

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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
This measure is specified at the program and plan level and there are currently no efforts to develop an eMeasure (eCQM).

Note for 3b3: Our understanding is that the Feasibility Score Card is only for eMeasures; consequently, we have not submitted this. Feasibility criteria were met during the initial endorsement review.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
This measure relies on standard data elements in administrative claims data (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes. A key advantage of using administrative claims data is that the time and cost of data collection for performance measurement purposes are relatively low because these data are already collected for other purposes.

Initial feasibility assessments were conducted using the RAND-UCLA modified Delphi process to rate the measure concepts with feasibility as one component of the assessment. On a 1-9 point scale, the measure concept of "periodic or comprehensive examination" was rated as an 8 or "definitely feasible" by the expert panel. During the empirical testing phase, our testing found that the critical data elements had missing/invalid data of <1% (Data 3c.1.), meeting or exceeding the guidance from the Centers

for Medicare and Medicaid Services regarding acceptable error rates. During measure development and testing, the measure specifications were made available through a publicly accessible website for public comment with additional broad email dissemination to a wide range of stakeholders. No concerns regarding feasibility were raised during this process.

Citation: Centers for Medicare & Medicaid Services. Medicaid and CHIP Statistical Information System (MSIS) File Specifications and Data Dictionary. 2010; http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MSIS/downloads/msisdd2010.pdf. Accessed August 10, 2013.

Data 3c.1 Percentage of Missing and Invalid Values for Critical Data Elements

PROGRAM 1
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.01%
Rendering Provider ID: 0.28%

PROGRAM 2
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.00%
Rendering Provider ID: 0.00%

PROGRAM 3
Member ID: 0.27%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.28%
Date of Service: 0.00%
Rendering Provider ID: 0.18%

PROGRAM 4
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.01%
Date of Service: 0.00%
Rendering Provider ID: 0.61%

PROGRAM 5
Member ID: 0.43%
Date of Birth: 0.02%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.00%
Rendering Provider ID: 0.67%

Maintenance of endorsement update: There have been no reports of feasibility issues with implementing this measure. Please see Use and Usability section.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.
This measure is intended to be transparent and available for widespread adoption. As such, it was purposefully designed to avoid using software or other proprietary materials that would require licensing fees. The measure specifications, including a

companion User Guide, are accessible through a website and can be used free of charge for non-commercial purposes. The main requirements of users is to ensure the quality of their source data and expertise to program the measures within their information systems, following the clear and detailed specifications. Technical assistance is available to users.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting Texas Health and Human Services Pay For Quality Program (Medicaid/CHIP) https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf Payment Program Texas Health and Human Services Pay For Quality Program (Medicaid/CHIP) https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf Quality Improvement (external benchmarking to organizations) Covered California http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf Quality Improvement (Internal to the specific organization) State Medicaid Agencies http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/? Michigan Healthy Kids Dental https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=00711 7B0011386&parentUrl=activeBids

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

1. Program and Sponsor: Texas Health and Human Services Commission

https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf

Purpose: Payment Program and Public Reporting

This measure has been adopted by the Texas Health and Human Services Commission as part of the Texas CHIP and Medicaid Dental Services Pay-for-Quality (P4Q) program. [Texas HHSC Uniform Managed Care Manual, Chapters 6.2.15. Effective Date 09/01/2017, Version 2.0].

This measure was also present in earlier iterations of the Texas Medicaid and CHIP quality programs since initial endorsement. We are referencing current use for this update.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies to the state of Texas CHIP and Medicaid programs (statewide application). There are two dental plans (i.e., the accountable entities) that serve Texas CHIP and Medicaid. In June 2017, there were 3,359,770 children enrolled in Texas Medicaid and CHIP (https://hhs.texas.gov/about-hhs/records-statistics/data-statistics/healthcare-statistics).

Level of Measurement and Setting: The measure is implemented at the plan and program level within the Texas Medicaid and CHIP programs.

2. Covered California, the California Health Benefit Exchange

http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf
http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-QDP-Issuer-Contract-and-Attachments.pdf

Purpose: Quality Improvement

This measure is included in the Covered California Qualified Health Plan Issuer Contract for 2017-019 for the Individual Market and the Covered California Qualified Dental Plan Issuer Contract for 2017-2019. The measure is to be reported annually.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies statewide. In March 2017 there were 85,000 enrollees 0-18 years old in 11 CC health plans (which may offer dental benefits and would therefore report on the dental quality measures). There were 5,100 children enrolled in 7 Qualified Dental Plans. (http://hbex.coveredca.com/data-research/)

Level of Measurement and Setting. The measure is implemented at the plan level with the Covered California program.

3. State Medicaid Agencies

http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

(Note: To access the data, a public user account must be created. We can help facilitate access to the data if needed.)

Purpose: Quality Improvement

The Medicaid|Medicare|CHIP Services Dental Association conducts an annual survey of state Medicaid programs and collects data specifically on which programs report Dental Quality Alliance measures.

In its 2015 profile (the most recent available), 10 states reported that they currently use this measure in their Medicaid and/or CHIP programs.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The 10 states are: Alabama, California, Florida, Idaho, Louisiana, Nevada, Oklahoma, Rhode Island, South Carolina, and West Virginia. Data are not provided on the number of accountable entities included.

4. Michigan Healthy Kids Dental Program

https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids

Note: Select Schedule A Work Statement link under File Attachments

Purpose: Quality Improvement

The Michigan Healthy Kids Dental Program has included this measure in the set of measures included in its Performance Monitoring Standards, which is currently included in the Request for Proposals and will be included in the contracts between the contracted dental plans and the State of Michigan.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The Healthy Kids Dental Program covers children enrolled in Michigan's Medicaid program statewide. The state intends to award two contracts. There are approximately 955,000 enrollees served by the Healthy Kids Dental Program.

Additional Information:
This measure was one of ten performance measures that focused on Dental Caries Prevention and Disease Management among children approved by the DQA. 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)

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable.

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not applicable.

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Per the annual survey conducted by the Medicaid|Medicare|CHIP Services Dental Association (MSDA), 10 Medicaid/CHIP agencies are implementing this measure for internal quality improvement. The measure is part of measure set included in the Request for Proposals (RFP) released by the Michigan Healthy Kids Dental Program. This measure is included in the Pay-For-Quality program and is publicly reported in the Texas Medicaid and CHIP programs. Additionally, this measure is a requirement for the Qualified Dental Plans to report to the Covered California, the state-based marketplace in California.

The DQA provides technical assistance to these and other users of DQA measures through webinars, resource document development, and one-on-one staff support. The DQA has an Implementation Committee dedicated to developing implementation and improvement resources.

In order to ensure transparency, incorporate learnings from implementation, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these

measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting. Implementation has focused on addressing questions related to how to use the measures in the context of broader quality improvement and clarifying questions related to the specifications.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

The DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

The DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.3. **Summarize the feedback obtained from other users**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures either have only limited baseline scores or will start reporting measures within the next year.

We are only aware of repeat measurements within the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after it was approved by the Dental Quality Alliance and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 2698361, 67.35, 69.23, 65.39
2015, 2929975, 69.12, 71.21, 66.49

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 452976, 59.43, 62.90, 58.23
2015, 341937, 63.41, 68.79, 63.62

These data suggest a trend in improvement over time. However, as noted above, these are initial performance data for one program. Most measure users are just now getting their quality measurement programs underway.

4b2. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
No unintended or negative consequences have been identified.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** NQF_Submission_OralEval_Appendix.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Dental Association on behalf of the Dental Quality Alliance
Co.2 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-
Co.3 **Measure Developer if different from Measure Steward:** American Dental Association on behalf of the Dental Quality Alliance
Co.4 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This project is headed by the DQA through its Measure Development and Maintenance Committee (formerly Research and Development Committee). The following individuals were responsible for executing and overseeing the maintenance process.

- Craig W. Amundson, DDS, General Dentist, HealthPartners, National Association of Dental Plans. Dr. Amundson serves as chair for the Committee.

- Mark Casey, DDS, MPH, Dental Director, North Carolina Department of Health and Human Services Division of Medical Assistance
- Natalia Chalmers, DDS, PhD, Diplomate, American Board of Pediatric Dentistry, Director, Analytics and Publication, DentaQuest Institute
- Frederick Eichmiller, DDS, Vice President & Science Officer, Delta Dental of Wisconsin
- Chris Farrell, RDH, BSDH, MPA, Oral Health Program Director, Michigan Department of Health and Human Services

This group is responsible for the maintenance of these measures and was also involved in the development and validation of the measure. All work of this Committee was distributed for review and formal vote and approval by the entire Dental Quality Alliance. (http://ada.org/dqa) The DQA is made up of representatives from 38 stakeholder organizations that represent all facets of the delivery system.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2013

Ad.3 **Month and Year of most recent revision:** 01, 2017

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 01, 2018

Ad.7 **Disclaimers:** Dental Quality Alliance Measures (Measures) and related data specifications, developed by the Dental Quality Alliance (DQA), are intended to facilitate quality improvement activities. These Measures are intended to assist stakeholders in enhancing quality of care. These performance Measures are not clinical guidelines and do not establish a standard of care. The DQA has not tested its Measures for all potential applications.

Measures are subject to review and may be revised or rescinded at any time by the DQA. The Measures may not be altered without the prior written approval of the DQA. The DQA shall be acknowledged as the measure steward in any and all references to the measure.

Measures developed by the DQA, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and DQA. Neither the DQA nor its members shall be responsible for any use of these Measures.

THE MEASURES ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND

Limited proprietary coding is contained in the Measure specifications for convenience.

For Proprietary Codes:

The code on Dental Procedures and Nomenclature is published in Current Dental Terminology (CDT), Copyright © 2017 American Dental

Association (ADA). All rights reserved.

This material contains National Uniform Claim Committee (NUCC) Health Care Provider Taxonomy codes (http://www.nucc.org/index.php?option=com_content&view=article&id=14&Itemid=125). Copyright © 2017 American Medical Association. All rights reserved.

Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. The DQA, American Dental Association (ADA), and its members disclaim all liability for use or accuracy of any terminologies or other coding contained in the specifications.

THE SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** In 2008, the Centers for Medicare and Medicaid Services (CMS) asked the ADA to lead the development of a broad coalition of organizations that would lead dentistry to improve the oral health of Americans through quality measurement and quality improvement. The ADA subsequently established the DQA. The DQA is a multi-stakeholder alliance comprised of 38 stakeholders (with organizations as members) from across the oral health community, including federal agencies, third-party payers, professional associations, and an individual member from the general public. The DQA's mission is to advance the field of performance measurement to improve oral health, patient care, and safety through a consensus building process.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2528
Measure Title: Prevention: Topical Fluoride for Children at Elevated Caries Risk, Dental Services
Measure Steward: American Dental Association on behalf of the Dental Quality Alliance
Brief Description of Measure: Percentage of enrolled children aged 1-21 years who are at "elevated" risk (i.e., "moderate" or "high") who received at least 2 topical fluoride applications within the reporting year.
Developer Rationale: Inequalities in oral health status and inadequate use of oral health care services are well documented. Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, Li, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009).

Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions. Evidence suggests that topical fluoride applied to children starting as early as six months of age is beneficial in preventing dental caries (Weyant et al. 2013). However, approximately three quarters of children younger than age 6 years did not have at least one visit to a dentist in the previous year (Edelstein & Chinn 2009). Evidence-based clinical recommendations suggest that topical fluoride should be applied at least every three to six months in children at elevated risk for caries (Weyant et al. 2013).

The proposed measure, Topical Fluoride for Children at Elevated Caries Risk – Dental Services, captures whether children at moderate or high caries risk received at least two topical fluoride applications as dental services. Because topical fluoride is indicated at 3-6 month intervals (2-4 times per year) for children at elevated caries risk, at least two applications are indicated during the reporting year. This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (professionally applied topical fluoride), including the frequency required for clinical effectiveness (at least every three-six months). Topical Fluoride allows plans and programs to assess whether children at risk for caries are receiving evidence-based preventive services and target performance improvement initiatives accordingly.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

Numerator Statement: Unduplicated number of enrolled children aged 1-21 years who are at "elevated" risk (i.e., "moderate" or "high") who received at least 2 topical fluoride applications as a dental service
Denominator Statement: Unduplicated number of enrolled children aged 1-21 years who are at "elevated" risk (i.e., "moderate" or "high")
Denominator Exclusions: Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Sep 18, 2014 **Most Recent Endorsement Date:** Sep 18, 2014

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- This measure assesses the percentage of children at moderate to high risk for caries who received at least two topical fluoride applications as dental services during the reporting year. Evidenced-based clinical guidelines recommends the specific topical fluoride agents for people who are at elevated risk of developing dental caries. (Weyant et al. 2013, full report, 0. 10)
- For children at elevated risk of developing caries specifically, the guidelines recommend applying 2.26 percent fluoride varnish at least every three to six months for children younger than 6 years old. For children 6-18 years old, the guidelines recommend 2.26 percent fluoride varnish at least every three to six months or 1.23 percent fluoride (APF).
    - 71 studies were included in evidence reviews, representing 82 citations. All studies included were controlled clinical trials.
    - This evidence received a grade of moderate by an expert panel, which is second on a three-point scale and denotes that evidence statements "are based on preliminary determination from the current best available evidence, but confidence in the estimate is constrained by one or more factors, such as: the number, size, or risk of bias of individual studies; inconsistency of findings across individual studies; limited applicability due to the populations of interest; or lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change could be large enough to alter the conclusion."
    - The clinical recommendations for fluoride among children and adolescents received an evidence grade of "in favor", which is the second highest out of six grading categories.

Citation:

Weyant RJ, Tracy SL, Anselmo TT, Beltrán-Aguilar ED, et al; American Dental Association Council on Scientific Affairs Expert Panel on Topical Fluoride Caries Preventive Agents. Topical fluoride for caries prevention: full report of the updated clinical recommendations and supporting systematic review. Available at: http://ebd.ada.org/contentdocs/Topical_fluoride_for_caries_prevention_2013_update_-_full_manuscript.pdf

Changes to evidence from last review

- ☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
- ☐ **The developer provided updated evidence for this measure:**

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3)→ QQC presented (Box 4)→ Contains Quantity: High (71 studies, 82 citations) Quality: Moderate, Consistency: High (Box 5b) → Rate as MODERATE

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer used data from four sources and referred to "program" level information and "plan" level information (Texas Medicaid, Florida CHIP, and Florida Medicaid programs, as well as national commercial data from Dental Service of Massachusetts, Inc.). The developer presented the total number of children enrolled in each program/plan. In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Florida CHIP, (3) Commercial Data, and (4) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011.
- The measure testing findings are consistent with other data indicating that children have sub-optimal utilization of dental services in general and preventive dental services in particular.
- The data source and sample size are sufficient to assess gaps in performance. The performance range of **18% to 35% in CY 2010** (year in which data were available for all five programs) indicates variation in topical fluoride application across programs.
- The developer did not provide more recent performance data, stating that due to the start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting.

Disparities

- Disparities were detected for age, geographic location, and race/ethnicity for all programs. The developer also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), Medicaid program type, CHIP dental plan, commercial product line, and preferred language for program communications. The developer detected disparities by income, health status, CHIP plan, and Medicaid program type, but data on all of these characteristics were not consistently available for all programs and present disparities data on those characteristics that were most consistently available and had the greatest standardization (i.e. race/ethnicity and geographic location).
- The developer provided an overview of the literature documenting the disparities in dental service use among children by age, race/ethnicity, and geographic region, including within vulnerable populations, much of which is summarized in three major national reports on oral health: the Surgeon General's report on Oral Health in

America in 2000, the IOM report, *Improving Access to Oral Health Care for Vulnerable and Underserved Populations,* and the IOM report, *Advancing Oral Health in America*.

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluation of Reliability and Validity (and composite construction, if applicable):
Staff evaluation of Scientific Acceptability

Questions for the Committee regarding reliability:

- *The staff is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *The staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure relies on **standard data elements in administrative claims data** (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes.

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- Texas Health and Human Services Commission: Texas Medicaid and CHIP: https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- The developer reports there has been no feedback indicating any significant issues related to the clarity or feasibility of implementing the measure specifications.

Additional Feedback:

- This measure was one of 10 performance measures approved by the Dental Quality Alliance (DQA) that focused on Dental Caries Prevention and Disease Management among children. The DQA was formed at the request of the 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 the 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."

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- The developer provides initial reporting data available from the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after approval by the DQA, but before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 1090952, 39.97, 41.57, 37.62
2015, 1334887, 41.75, 44.70, 38.15

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 108704, 33.01, 35.45, 32.99
2015, 79693, 37.50, 41.44, 37.71

- The developer noted that these data suggest a trend in improvement over time. These are initial performance data for one program, however, since most measure users are just now getting their quality measurement programs underway.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer reports no unintended or negative consequences have been identified.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures
Related or competing measures • N/A **Harmonization** • N/A
Committee pre-evaluation comments **Criterion 5: Related and Competing Measures**

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Scientific Acceptability

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?
   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
   ☐Yes (go to Question #5)
   ☒No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☐Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

6. **RATING (score level) -** What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
   Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☐High (go to Question #8)
   ☐Moderate (go to Question #8)
   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☒Yes (go to Question #9)
   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☒Yes (go to Question #10)
   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?
   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.
   ☒Yes (go to Question #2)
   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?
   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?
   ☐Yes (please explain below then go to Question #3)
   ☐No (go to Question #3)
   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?
   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*
   ☐Yes (please explain below then go to Question #4)
   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
   ☐Yes (please explain below then go to Question #5)
   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
   ☐Yes (please explain below then go to Question #6)
   ☐No (go to Question #6)

☒Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☐Yes (go to Question #11)
    ☒No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☐Yes (go to Question #12)
    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #14)
☐Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

13. Was other validity testing reported?

☒Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

TIPS: Prior validity studies of the same data elements may be submitted

☒Yes (go to Question #15)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #16)
☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Prevention: Topical Fluoride for Children at Elevated Caries Risk, Dental Services
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure title
Date of Submission: 2/11/2014

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☐ Health outcome: Click here to name the health outcome

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome
X☐ Process: Receipt of evidence-based preventive service - topical fluoride application - during the reporting period
☐ Structure: Click here to name the structure
☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not applicable.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

Not applicable.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Topical Fluoride for Children at Elevated Caries Risk indicates the percentage of children at moderate to high risk for caries who received at least two topical fluoride applications as dental services during the reporting year. Evidence suggests that topical fluoride applied to children starting as early as six months of age is beneficial in preventing dental caries (Weyant et al. 2013). Evidence-based clinical recommendations also suggest that topical fluoride should be applied at least every three to six months in children at elevated risk for caries (Weyant et al. 2013). This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (professionally applied topical fluoride), including the frequency required for clinical effectiveness (at least every three-six months). As described in 1b1 (Importance), dental caries is the most common chronic disease in children in the U.S. and a significant percentage of children have untreated dental caries. Dental decay causes significant short- and long-term adverse consequences for children's health and functioning. As detailed below, professionally applied topical fluoride has demonstrated effectiveness in reducing caries among children at elevated caries risk, thereby improving oral health, overall health, and overall well-being.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?
☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***
☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***
☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***
☐ Other – ***complete section 1a.8***
Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION
1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):
Full Report: Weyant RJ, Tracy SL, Anselmo TT, Beltrán-Aguilar ED, et al; American Dental Association Council on Scientific Affairs Expert Panel on Topical Fluoride Caries Preventive Agents. Topical fluoride for caries prevention: full report of the updated clinical recommendations and supporting systematic review. Available at: http://ebd.ada.org/contentdocs/Topical_fluoride_for_caries_prevention_2013_update_-_full_manuscript.pdf.

Condensed version: Weyant RJ, Tracy SL, Anselmo TT, Beltrán-Aguilar ED, et al; American Dental Association Council on Scientific Affairs Expert Panel on Topical Fluoride Caries Preventive Agents. J Am Dent Assoc. 2013 Nov;144(11):1279-91. Topical fluoride for caries prevention: executive summary of the updated clinical recommendations and supporting systematic review. Available at: http://ebd.ada.org/contentdocs/JADA_updated_executive_summary_Nov_2013.pdf.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.
Summary of Clinical Recommendations: "For people who are at an elevated risk of developing dental caries, the panel makes clinical recommendations for the use of specific topical fluoride agents; these recommendations are based on the evidence statements and the balance of benefits with potential harm. The panel recommends topical fluoride agents only for people who are at elevated risk of developing dental caries." (Weyant et al. 2013, full report, 0. 10)

Age –Specific Recommendations: "The panel recommends the following for people at risk of developing dental caries: 2.26% fluoride varnish or 1.23% fluoride (APF) gel, or a prescription-strength, home-use 0.5% fluoride gel or paste or 0.09% fluoride mouthrinse for 6 years or older. Only 2.26% fluoride varnish is recommended for children younger than 6 years. The strengths of the recommendations for the recommended products varied from "in favor" to "expert opinion for." As part of the evidence-based approach to care, these clinical recommendations should be integrated with the practitioner's professional judgment and the patient's needs and preferences." (Weyant et al. 2013, full report, p. 10)

For children at elevated risk of developing caries specifically, the recommendations are "in favor" for:
- "2.26 percent fluoride varnish at least every three to six months" for children younger than 6 years
- "2.26 percent fluoride varnish at least every three to six months OR 1.23 percent fluoride (APF) gel for four minutes at least every three to six months" for children 6-18 years

(Weyant et al., 2013, p. 11, Table 1)

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade: The grade for both bulleted items is **"in favor"** which is defined as: "Evidence favors providing this intervention." This is the second highest recommendation out of a six-point scale. The grading system was adapted from that used by the U.S. Preventive Services Task Force. (Weyant et al. 2013, full report, p. 11, Table 1; p. 20, Table 6)

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Strong: Evidence strongly supports providing this intervention.
In Favor: Evidence favors providing this intervention.

Weak: Evidence suggests implementing this intervention after alternatives have been considered.
Expert Opinion For:† Evidence is lacking; the level of certainty is low. Expert opinion guides this recommendation
Expert Opinion Against:† Evidence is lacking; the level of certainty is low. Expert opinion suggests not implementing this intervention.
Against: Evidence suggests not implementing this intervention or discontinuing ineffective procedures.

† The USPSTF system defines this category of evidence as "insufficient"; "grade I indicates that the evidence is insufficient to determine the relationship between benefits and harms (i.e., net benefit)." The corresponding recommendation grade "I" is defined as follows: "The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined."

Grades definitions can be found at Weyant et al. 2013, full report, p. 20, Table 6. The grading system was adapted from that used by the U.S. Preventive Services Task Force. (U.S. Preventive Services Task Force. Methods and processes. Available at: www.uspreventiveservicestaskforce.org/methods.htm.)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**
Same as that provided for the guidelines provided in 1a.4.1.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?
☐ **X**Yes → ***complete section 1a.7***
☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION
1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):
Not applicable.

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.
Not applicable.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade: Not applicable.
1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)
Not applicable.

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**
Not applicable.
Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE
1a.6.1. Citation (*including date*) and **URL** (*if available online*):
Not applicable.

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**
Not applicable.
Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The following three clinical questions were addressed:

- "In primary and permanent teeth, does the use of a topical fluoride compared to no topical fluoride reduce the incidence of new lesions, or arrest or reverse existing coronal and/or root caries?"
- "For primary and permanent teeth, is one topical fluoride agent more effective than another in reducing the incidence of, or arresting or reversing coronal and/or root caries?"
- "Does the use of prophylaxis before application of topical fluoride reduce the incidence of caries to a greater extent than the application of topical fluoride without prophylaxis?"
  (Weyant et al., 2013, full report, pp. 7-8)

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

For children at elevated risk of developing caries specifically, the recommendations are "in favor" for:

- "2.26 percent fluoride varnish at least every three to six months" for children younger than 6 years
- "2.26 percent fluoride varnish at least every three to six months OR 1.23 percent fluoride (APF) gel for four minutes at least every three to six months" for children 6-18 years

(Weyant et al., 2013, p. 11, Table 1)

Grade: The evidence grade for both bulleted items is **"moderate"** which is defined as: "This statement is based on preliminary determination from the current best available evidence, but confidence in the estimate is constrained by one or more factors, such as: the number, size, or risk of bias of individual studies; inconsistency of findings across individual studies; limited applicability due to the populations of interest; or lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change could be large enough to alter the conclusion." (Weyant et al., 2013, full report, pp. 18-19, Table 4)

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

High: This statement is strongly established by the best available evidence; the conclusion is unlikely to be affected strongly by the results of future studies. The body of evidence usually includes consistent results from well-designed, well-conducted studies in representative populations. This conclusion is unlikely to be strongly affected by the results of future studies.

Moderate: This statement is based on preliminary determination from the current best available evidence, but confidence in the estimate is constrained by one or more factors, such as: the number, size, or risk of bias of individual studies; inconsistency of findings across individual studies; limited applicability due to the populations of interest; or lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change could be large enough to alter the conclusion."

Low: The available evidence is insuffi cient to support the statement, or the statement is based on extrapolation from the best available evidence. Evidence is insufficient or the reliability of estimated effects is limited by factors such as: the limited number or size of studies; important flaws in study design or methods leading to high risk of bias; inconsistency of findings across individual studies; gaps in the chain of evidence; findings not applicable to the populations of interest; or a lack of information on important health outcomes. More information could allow a reliable estimation of effects on health outcomes.
(Weyant 2013, full report, pp. 18-19)

The grading system was adapted from that used by the U.S. Preventive Services Task Force (U.S. Preventive Services Task Force. Available at: www.uspreventiveservicestaskforce.org/methods.htm.)

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 1965-2012

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

71 studies included in evidence reviews, representing 82 citations. All studies included were controlled clinical trials.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The quality of the evidence was rated by the expert panel as "moderate" - i.e., the evidence statements "are based on preliminary determination from the current best available evidence, but confidence in the estimate is constrained by one or more factors, such as: the number, size, or risk of bias of individual studies; inconsistency of findings across individual studies; limited applicability due to the populations of interest; or lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change could be large enough to alter the conclusion."

However, the clinical recommendations for fluoride among children and adolescents received an evidence grade of "in favor", which is the second highest out of six grading categories. The expert panel not only made recommendations based on the study designs, but also on an evaluation on the *net benefit* of the interventions, explicitly balancing benefits to potential harms in conjunction with the level of the certainty of the evidence. The full methodology is provided in Weyant et al., full report, 2013.

The evidence directly pertains to both the measure focus and the measure target population.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Recommendations:

- "2.26 percent fluoride varnish at least every three to six months" for children younger than 6 years
- "2.26 percent fluoride varnish at least every three to six months OR 1.23 percent fluoride (APF) gel for four minutes at least every three to six months" for children 6-18 years

Estimates of Benefit in Support of Recommendations:

(1) 2.26% Fluoride Varnish

"The results of meta-analyses for primary teeth indicate tha the application of 2.26% fluoride varnish has a statistically significant effect (SMD -0.19 [95% CI: -0.31, -0.08)on caries prevention as measured by increment or incidence using surface-level data." Weyant et al., full report, 2013, p. 25

"The results of meta-analyses for permanent teeth indicate that 2.26% fluoride varnish has a statistically significant effect (SMD= -0.38 [95% CI: -0.53, -0.24])on caries prevention as measured by increment or incidence using surface-level data." Weyant et al., full report, 2013, p. 25

Evidence Profile (Weyant et al., full report, 2013, pp. 26-27):

(a) Primary teeth (children under age 6):

- Level of certainty: Moderate
- Benefit: Yes (smaller caries increment or incidence with topical fluoride use).
  - Standardized mean difference=-0.19 [-0.31, -0.08]
  - Prevented fraction=0.27
  - Number needed to treat for control rate of 1 DMFS per year = 4
- Adverse events or harms: Little potential for harms if swallowed
- Benefit-harm assessment (Net benefit rating): Benefits outweigh potential harms
- Strength of clinical recommendation: In favor

(b) Permanent teeth (children):
- Level of certainty: Moderate
- Benefit: Yes (smaller caries increment or incidence with topical fluoride use).
  - Standardized mean difference=-0.38 [-0.53, -0.24]
  - Prevented fraction=0.36
  - Number needed to treat for control rate of 1 DMFS per year = 3
- Adverse events or harms: None if used as manufacturers recommend
- Benefit-harm assessment (Net benefit rating): Benefits outweigh potential harms
- Strength of clinical recommendation: In favor

The table below (Table 8 from the report) summarizes the findings.

Table 8. Summary of standardized mean differences from meta-analysis and individual studies for 2.26% fluoride varnish studies

Outcome Measures	Number and type* of studies	Number of participants**	Standardized Mean Difference [95% Confidence Interval] (negative favors intervention, positive favors control)
Meta-analysis results: Primary teeth			
d(e/m)fs, increment or incidence †	6 RCT[31-33, 35-39] and 2 CCT[40, 41]	3,409**	**-0.19 [-0.31, -0.08]**
Meta-analysis results: Permanent teeth			
D(M)FS, increment or incidence†	8 RCT[31-33, 42-44, 46-49, 52] and 1 CCT[54]	2,574	**-0.38 [-0.53, -0.24]**
Root caries, meta-analysis results			
Root caries increment	2 RCT[50, 51]	132	**-0.67 [-1.14, -0.20]**
Individual study results			
Combined dentition	1 CCT[55]	390	**DMFS + dmfs: -1.47 [-1.70, -1.25]** **DMFT + dmft: -1.15 [-1.37, -0.94]**
DMFT	1 CCT[53]	77	**-0.13 [-0.58, 0.32]**
DS occlusal surfaces	1 RCT[45]	79	**-0.54 [-1.06, -0.03]**

Notes: * RCT = randomized controlled trial; CCT = controlled clinical trial (non-randomized); **Including all participants (not using cluster-adjusted number of participants or numbers of clusters); †all stages used if cavitated data not available; parentheses indicate that component was included in some of the combined results and not others

(2) 1.23% fluoride (APF) gel

"The panel concluded with moderate certainty that there is a benefit of APF gel (1.23% fluoride) application up to every three months for 4G minutes for caries prevention in the permanent teeth of 6-14 year olds. This statement is based on meta-analysis of 12 studies with moderate to high bias scores and including over 4,000 participants; although there was some inconsistency, there was low statistical heterogeneity (I2=43) between the studies." (Weyant, full report, 2013, p. 33)

Evidence Profile (Weyant et al., full report, 2013, p. 34):

Permanent teeth (children):
- Level of certainty: Moderate
- Benefit: Yes (smaller caries increment or incidence with topical fluoride use).
  - Standardized mean difference=-0.25 [-0.33, -0.16]

- Prevented fraction=0.27
- Number needed to treat for control rate of 1 DMFS per year = 4

- Adverse events or harms: None if used as manufacturers recommend
- Benefit-harm assessment (Net benefit rating): Benefits outweigh potential harms
- Strength of clinical recommendation: In favor

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Potential harms evaluated included: (1) nausea and vomiting associated with the ingestion of topical fluorides and (2) dental fluorosis while tooth enamel is developing until approximately age 6, due to daily ingestion of topical fluoride, such as from toothpaste or from prescription home gels.

"There is less of a concern with professionally-applied topical fluorides that have much longer intervals between applications [citing Wong et al. 2010]. Additionally, fluoride varnish has less potential for harms than other forms of high concentration topical fluoride because the amount of fluoride that is placed in the mouth with fluoride varnish is approximately one-tenth that of other professionally-applied products [citing Beltran-Aguilar et al. 2000]. The panel judged that the benefits outweighed the potential for harms for all professionally-applied or prescription-strength topical fluorides and age groups except for children under age 6, where the risk of swallowing and associated events (particularly nausea and vomiting) outweighed the potential benefits for all professionally-applied or prescription-strength topical fluorides except 2.26% fluoride varnish." (Weyant et al., 20130, p. 10)

Citations

Beltran-Aguilar ED, Goldstein JW, Lockwood SA. Fluoride varnishes - A review of their clinical use, cariostatic mechanism, efficacy and safety. JADA 2000;131(May):589-96.

Weyant RJ, Tracy SL, Anselmo TT, Beltrán-Aguilar ED, et al; American Dental Association Council on Scientific Affairs Expert Panel on Topical Fluoride Caries Preventive Agents. Topical fluoride for caries prevention: full report of the updated clinical recommendations and supporting systematic review. Available at: http://ebd.ada.org/contentdocs/Topical_fluoride_for_caries_prevention_2013_update_-_full_manuscript.pdf

Wong MC, Glenny AM, Tsang BW, et al. Topical fluoride as a cause of dental fluorosis in children. Cochrane Database of Systematic Reviews 2010;Jan 20(1).

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

This review was published on November 2013 and reflects the latest evidence.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

Not applicable.

1a.8.2. Provide the citation and summary for each piece of evidence.

Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
4_Evidence_fluoride.docx
1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?
Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.
No

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*
If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
Inequalities in oral health status and inadequate use of oral health care services are well documented. Dental caries is the most common chronic disease in children in the United States (NCHS 2012). In 2009–2010, 14% of children aged 3 –5 years had untreated dental caries. Among children aged 6–9 years, 17% had untreated dental caries, and among adolescents aged 13–15, 11% had untreated dental caries (Dye, Li, and Thorton-Evans 2012). Dental decay among children has significant short- and long-term adverse consequences (Tinanoff and Reisine 2009). Childhood caries is associated with increased risk of future caries (Gray, Marchment, and Anderson 1991; O´Sullivan and Tinanoff 1996; Reisine, Litt, and Tinanoff 1994), missed school days (Gift, Reisine, and Larach 1992; Hollister and Weintraub 1993), hospitalization and emergency room visits (Griffin et al. 2000; Sheller, Williams, and Lombardi 1997) and, in rare cases, death (Casamassimo et al. 2009).

Identifying caries early is important to reverse the disease process, prevent progression of caries, and reduce incidence of future lesions. Evidence suggests that topical fluoride applied to children starting as early as six months of age is beneficial in preventing dental caries (Weyant et al. 2013). However, approximately three quarters of children younger than age 6 years did not have at least one visit to a dentist in the previous year (Edelstein & Chinn 2009). Evidence-based clinical recommendations suggest that topical fluoride should be applied at least every three to six months in children at elevated risk for caries (Weyant et al. 2013).

The proposed measure, Topical Fluoride for Children at Elevated Caries Risk – Dental Services, captures whether children at moderate or high caries risk received at least two topical fluoride applications as dental services. Because topical fluoride is indicated at 3-6 month intervals (2-4 times per year) for children at elevated caries risk, at least two applications are indicated during the reporting year. This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (professionally applied topical fluoride), including the frequency required for clinical effectiveness (at least every three-six months). Topical Fluoride allows plans and programs to assess whether children at risk for caries are receiving evidence-based preventive services and target performance improvement initiatives accordingly.

Note: Procedure codes contained within claims data are the most feasible and reliable data elements for quality metrics in dentistry, particularly for developing programmatic process measures to assess the quality of care provided by programs (e.g., Medicaid, CHIP) and health/dental plans. In dentistry, diagnostic codes are not commonly reported and collected, precluding direct outcomes assessments. Although some programs are starting to implement policies to capture diagnostic information, evidence-based process measures are the most feasible and reliable quality measures at programmatic and plan levels at this point in time.

[Complete citations provided in 1c4 and in Evidence Submission Form.]

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Below are the testing data and results that met scientific acceptability criteria for endorsement. Because there were no changes in the data source, level of analysis or setting, additional testing has not been conducted.

Data Sources:

We used data from four sources and refer to "program" level information and "plan" level information. We included data for publicly insured children in the Texas Medicaid, 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 four programs collectively represent different delivery system models. The Texas Medicaid data represented dental fee-for-service. The Florida CHIP data included da ta 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. Data from calendar years 2010 and 2011 were used for all programs except Florida Medicaid. Full-year data for CY 2011 were not available for Florida Medicaid. Therefore, we report only CY 2010 data for Florida Medicaid.

In the data summaries, "Programs" refer to population data from (1) Texas Medicaid, (2) Florida CHIP, (3) Commercial Data, and (4) Florida Medicaid. "Plans" refer to data from the two dental plans that served Florida CHIP members in both 2010 and 2011.

Below we provide summary data for each of the four programs and two plans individually.

Programs

Our source data for the testing included children 0-20 years in each program. The numbers of children ages 0-20 years enrolled at least one month in each program were as follows:

Texas Medicaid, 2011: 3,544,247
Texas Medicaid, 2010: 3,393,963
Florida CHIP, 2011: 317,146
Florida CHIP, 2010: 315,975
Commercial, 2011: 184,152
Commercial, 2010: 189,968
Florida Medicaid, 2010: 2,068,670

Within these programs, we had claims data available in both years for two dental managed care plans in Florida CHIP. We also report rates for those two plans separately.

Plan 1, 2010: 77,255
Plan 2, 2010: 116,388
Plan 1, 2011: 140,986
Plan 2, 2011: 168,191

Data 1b.2. Performance Scores for Topical Fluoride, Dental Services

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

Program 1, CY 2011:	37.13%	(	0.3713	,	0.0004	,	0.3704	,	0.3722	)
Program 2, CY 2011:	27.15%	(	0.2715	,	0.0020	,	0.2676	,	0.2754	)
Program 3, CY 2011:	22.04%	(	0.2204	,	0.0020	,	0.2165	,	0.2243	)
Program 1, CY 2010:	34.96%	(	0.3496	,	0.0005	,	0.3487	,	0.3505	)
Program 2, CY 2010:	22.63%	(	0.2263	,	0.0019	,	0.2225	,	0.2301	)
Program 3, CY 2010:	35.04%	(	0.3504	,	0.0023	,	0.3458	,	0.3550	)
Program 4, CY 2010:	18.16%	(	0.1816	,	0.0009	,	0.1799	,	0.1833	)
Plan 1, CY 2011:	25.50%	(	0.2550	,	0.0030	,	0.2491	,	0.2609	)
Plan 2, CY 2011:	28.69%	(	0.2869	,	0.0027	,	0.2815	,	0.2923	)
Plan 1, CY 2010:	23.24%	(	0.2324	,	0.0048	,	0.2230	,	0.2418	)
Plan 2, CY 2010 :	23.76%	(	0.2376	,	0.0034	,	0.2309	,	0.2443	)

The measure score range of 18% to 35% in CY 2010 (year in which data were available for all four programs) indicates a significant performance gap overall. Two-thirds or more of children identified as being at elevated risk for caries do not receive the

evidence-based recommendations of at least two topical fluoride applications during the reporting year. In addition, these results demonstrate the ability of the measure to identify variations in performance between programs.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

The measure testing findings are consistent with other data indicating that children have sub-optimal utilization of dental services in general and preventive dental services in particular. Although comprehensive dental benefits are covered under Medicaid and the Children's Health Insurance Program (CHIP), there are significant variations in use of dental services overall across states, ranging from approximately 25% to 69% (CMS EPSDT Data, FY 2011). Similar variation between states is observed among children 0-20 years of age enrolled in commercial dental plans (ADA 2013). With respect to preventive dental services more specifically, 14% to 58% of children enrolled in Medicaid/CHIP for at least 90 continuous days receive any preventive dental services (CMS EPSDT Data, FY 2011). Even among the highest performing states, 42% of publicly-insured children do not receive any type of preventive dental service during the year.

[Complete citations provided in 1c4 and in Evidence Submission Form Template.]

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The same data sources were used as described in 1b.2. The data below summarizes performance data by age, geographic location, and race/ethnicity for CY 2011 (CY 2010 for one program) with the p-values from chi-square tests used to detect whether there were statistically significant differences in performance between groups. The results demonstrate that there are disparities by age, geographic location and race/ethnicity. In addition, we also evaluated whether the measure could detect disparities by income (within program), children's health status (based on their medical diagnoses), Medicaid program type, CHIP dental plan, commercial product line, and preferred language for program communications. We additionally detected disparities by income, health status, CHIP plan, and Medicaid program type, but data on all of these characteristics were not consistently available for all programs so we are presenting disparities data on those characteristics that were most consistently available and had the greatest standardization

Data1b.4. Disparities in Performance by Child Age, Geographic Location and Race/Ethnicity
PROGRAM 1
Overall performance score: 37.13%
Scores by Age
Age 1-2 years: 6.21%
Age 3-5 years: 43.07%
Age 6-7 years: 43.64%
Age 8-9 years: 42.03%
Age 10-11 years: 40.50%
Age 12-14 years: 34.83%
Age 15-18 years: 24.93%
Age 19-20 years: 11.75%
p-value from Chi-square test: <0.0001
Scores by Geographic Location
Urban: 37.87%
Rural: 32.50%
p-value from Chi-square test: <0.0001
Scores by Race
Non-Hispanic White: 30.37%
Non-Hispanic Black: 29.68%
Hispanic: 40.84%
p-value from Chi-square test: <0.0001

PROGRAM 2
Overall performance score: 27.15%
Scores by Age
Age 1-2 years: n/a
Age 3-5 years: 30.00%
Age 6-7 years: 37.81%
Age 8-9 years: 34.88%
Age 10-11 years: 31.60%
Age 12-14 years: 27.14%
Age 15-18 years: 18.60%
Age 19-20 years: n/a
p-value from Chi-square test: <0.0001
Scores by Geographic Location
Urban: 26.96%
Rural: 30.64%
p-value from Chi-square test: <0.0001
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 3
Overall performance score: 22.04%
Scores by Age
Age 1-2 years: 25.93%
Age 3-5 years: 34.24%
Age 6-7 years: 34.11%
Age 8-9 years: 33.97%
Age 10-11 years: 32.26%
Age 12-14 years: 28.78%
Age 15-18 years: 15.08%
Age 19-20 years: 2.22%
p-value from Chi-square test: <0.0001
Scores by Geographic Location
Urban: 22.13%
Rural: 19.71%
p-value from Chi-square test: 0.025
Scores by Race
Non-Hispanic White: n/a
Non-Hispanic Black: n/a
Hispanic: n/a
p-value from Chi-square test n/a

PROGRAM 4
Overall performance score: 18.16%
Scores by Age
Age 1-2 years: 17.17%
Age 3-5 years: 21.43%
Age 6-7 years: 21.19%
Age 8-9 years: 21.44%
Age 10-11 years: 19.47%
Age 12-14 years: 16.86%
Age 15-18 years: 12.53%
Age 19-20 years: 7.45%
p-value from Chi-square test: <0.0001
Scores by Geographic Location

Urban: 18.16%
Rural: 17.32%
p-value from Chi-square test: 0.025
Scores by Race
Non-Hispanic White: 21.64%
Non-Hispanic Black: 15.02%
Hispanic: 17.74%
p-value from Chi-square test: <0.0001

Note: N/A for age indicates that those ages are not within the program's age eligibility. N/A for race/ethnicity indicates that those programs did not collect race/ethnicity data or had high rates of missing data

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

There is extensive literature documenting disparities in dental service use among children by age, race/ethnicity, and geographic region, including within vulnerable populations, much of which is summarized in three major national reports on oral health: the Surgeon General's report on Oral Health in America in 2000, the IOM report, Improving Access to Oral Health Care for Vulnerable and Underserved Populations, and the IOM report, Advancing Oral Health in America.

With respect to preventive dental services, there are documented disparities. Using data from the National Survey of Children's Health, Edelstein and Chinn (2009) noted disparities in access to preventive dental services by race and income: "Stepwise disparities in access to preventive dental services are evident by race and income in ways that parallel Medical Expenditure Panel Survey findings. White parents report higher use of preventive dental services than do black or Hispanic parents (77%, 66%, and 61%, respectively). Poor parents report less use of services than do low income, middle class, and higher-income parents (58%, 66%, 77%, and 82%, respectively)" (Edelstein & Chinn, 2009, p.418). A recent analysis by Bouchery (2013) of the Medicaid Analytic eXtract files for nine states found variations in the percentage of children receiving a preventive dental visit by age, race and ethnicity, and geographic area. Specifically, relative to the reference group of 9 year olds, the percentage point change in the probability of having a dental preventive services was -27.6 for 3 years old; -8.6 for 6 years, -2.2 for 12 years and -15.4 for 15 years (all significant at $p<0.0001$); relative to the reference group of white, non-Hispanic, the percentage point change was -1.8 for black non-Hispanic and 7.8 for Hispanic ($p<0.0001$ for both); relative to the reference group of small metro area, the percentage point change was 5.9 for large metro area ($p<0.0001$).

Sources

Bouchery, E. 2013. "Utilization of Dental Services among Medicaid-Enrolled Children." Medicare & Medicaid Research Review. 3(3) E1-16. Available at: https://www.cms.gov/mmrr/Downloads/MMRR2013_003_03_b04.pdf.

Dietrich, T., C. Culler, R. Garcia, and M. M. Henshaw. 2008. Racial and ethnic disparities in children's oral health: The National Survey of Children's Health. Journal of the American Dental Association 139(11):1507-1517.

Dye BA, Li X, Thorton-Evans G. Oral health disparities as determined by selected healthy people 2020 oral health objectives for the United States, 2009-2010. NCHS Data Brief 2012(104):1-8.U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research.

Edelstein, B. L. and C. H. Chinn. 2009. "Update on Disparities in Oral Health and Access to Dental Care for America´s Children." Acad Pediatr 9(6): 415-9.

Institute of Medicine (U.S.). Committee on an Oral Health Initiative. Advancing oral health in America. Washington, D.C.: National Academies Press; 2011.

Institute of Medicine and National Research Council. Improving access to oral health care for vulnerable and underserved populations. Washington, D.C.: National Academies Press; 2011.

Kenney, G. M., J. R. McFeeters, and J. Y. Yee. 2005. Preventive dental care and unmet dental needs among low-income children. American Journal of Public Health 95(8):1360-1366.

Lewis, C., W. Mouradian, R. Slayton, and A. Williams. 2007. Dental insurance and its impact on preventative dental care visits for U.S. children. Journal of the American Dental Association 138(3):369-380.

U.S. Dept. of Health and Human Services, National Institute of Dental and Craniofacial Research. Oral health in America : a report of the Surgeon General. Rockville, Md.: U.S. Public Health Service, Dept. of Health and Human Services; 2000.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Ears, Nose, Throat (ENT)

De.6. **Non-Condition Specific***(check all the areas that apply):*
Access to Care, Disparities Sensitive, Health and Functional Status : Change, Health and Functional Status : Total Health, Primary Prevention

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, Populations at Risk

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.ada.org/~/media/ADA/Science%20and%20Research/Files/DQA_2018_Dental_Services_Topical_Fluoride.pdf?la=en

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
1. No changes to the measure specifications

2. Measure specification website updated to be more user friendly

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Unduplicated number of enrolled children aged 1-21 years who are at "elevated" risk (i.e., "moderate" or "high") who received at least 2 topical fluoride applications as a dental service

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see section S14.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Unduplicated number of enrolled children aged 1-21 years who are at "elevated" risk (i.e., "moderate" or "high")

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Please see Section S14.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Medicaid/CHIP programs should exclude those individuals who do not qualify for dental benefits. The exclusion criteria should be reported along with the number and percentage of members excluded.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
There are no other exclusions than those described above

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
This measure is stratified by age using the following categories:

1-2; 3-5; 6-7; 8-9; 10-11; 12-14; 15-18; 19-20

No new data are needed for this stratification. Please see attached specifications for complete measure details.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

Topical Fluoride Intensity Calculation for Children at Elevated Caries Risk

1. Use administrative enrollment and claims data for a single year. When using claims data to determine service receipt, include both paid and unpaid claims (including pending, suspended, and denied claims).

2. Check if the enrollee meets age criteria at the last day of the reporting year:
a. If child is >=1 and < 21, then proceed to next step.
b. If age criteria are not met or there are missing or invalid field codes (e.g., date of birth), then STOP processing. This enrollee does not get counted.

3. Check if subject is continuously enrolled for the reporting year (12 months) with a gap of no more than 31 days (one month gap for programs that determine eligibility on a monthly basis):

a. If subject meets continuous enrollment criterion, then proceed to next step.
b. If subject does not meet enrollment criterion, then STOP processing. This enrollee does not get counted.

YOU NOW HAVE THE COUNT OF THOSE WHO MEET THE AGE AND ENROLLMENT CRITERIA

4. Check if subject is at "elevated risk":
a. If subject meets ANY of the following criteria, then include in denominator:
i. the subject has a CDT Code among those in Table 1 in the reporting year,
OR
ii. the subject has a CDT Code among those in Table 1 in any of the three years prior to the reporting year, (NOTE: The subject does not need to be enrolled in any of the prior three years for the denominator enrollment criteria; this is a "look back" for enrollees who do have claims experience in any of the prior three years.)
OR
iii. the subject has a visit with a CDT code = (D0602 or D0603) in the reporting year.
b. If the subject does not meet any of the above criteria for elevated risk, then STOP processing. This enrollee will not be included in the measure denominator.

YOU NOW HAVE THE DENOMINATOR (DEN): Enrollees who are at "elevated risk"

5. Check if subject received at least two fluoride applications as dental service during the reporting year – at least two unique dates of service when topical fluoride was provided. Service provided on each date of service should satisfy the following criteria:
a. If [CDT CODE] = D1206 or D1208 , and
b. If [RENDERING PROVIDER TAXONOMY] code = any of the NUCC maintained Provider Taxonomy Codes in Table 1 below, then include in numerator; proceed to next step.
c. If both a AND b are not met, then the service was not a "dental service"; STOP processing. This enrollee is already included in the denominator but will not be included in the numerator.

Note 1: No more than one fluoride application can be counted for the same member on the same date of service.
Note 2: All claims with missing or invalid CDT CODE, missing or invalid NUCC maintained Provider Taxonomy Codes, or NUCC maintained Provider Taxonomy Codes that do not appear in Table 2 should not be included in the numerator.

YOU NOW HAVE NUMERATOR (NUM) COUNT: Enrollees at "elevated risk" who received fluoride as a dental service

6. Report
a. Unduplicated number of enrollees in numerator
b. Unduplicated number of enrollees in denominator
c. Measure Rate (NUM/DEN)
d. Rate stratified by age

Table 1: CDT Codes to identify "elevated risk"

D2140	D2394	D2630	D2720	D2791	D3120
D2150	D2410	D2642	D2721	D2792	D3220
D2160	D2420	D2643	D2722	D2794	D3221
D2161	D2430	D2644	D2740	D2799	D3222
D2330	D2510	D2650	D2750	D2930	D3230
D2331	D2520	D2651	D2751	D2931	D3240
D2332	D2530	D2652	D2752	D2932	D3310
D2335	D2542	D2662	D2780	D2933	D3320
D2390	D2543	D2663	D2781	D2934	D3330
D2391	D2544	D2664	D2782	D2940	D2941
D2392	D2610	D2710	D2783	D2950	D1354
D2393	D2620	D2712	D2790	D3110	

Table 2: NUCC maintained Provider Taxonomy Codes classified as "Dental Service"*

122300000X	1223P0106X	1223X0008X	261QF0400X
1223D0001X	1223P0221X	1223X0400X	261QR1300X
1223D0004X	1223P0300X	124Q00000X+	125Q00000X
1223E0200X	1223P0700X	125J00000X	
1223G0001X	1223S0112X	125K00000X	

*Services provided by County Health Department dental clinics may also be included as "dental" services.
+Only dental hygienists who provide services under the supervision of a dentist should be classified as "dental" services. Services provided by independently practicing dental hygienists should be classified as "oral health" services and are not applicable for this measure.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
Not applicable.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
Not applicable.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
Not applicable.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Outpatient Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable.

2. Validity – See attached Measure Testing Submission Form
5_Testing_top_flouride.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
No

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Prevention: Topical Fluoride for Children at Elevated Caries Risk, Dental Services
Date of Submission: 2/12/2014
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.

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] 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] 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]

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]

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; [14,15] 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] **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
☒ 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:	☐ other:

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 four sources and refer to "program" level information and "plan" level information. We included data for publicly insured children in the Texas Medicaid, 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 four programs collectively represent different delivery system models. The Texas Medicaid data represented dental fee-for-service. 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 <u>all</u> 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 <u>measured entities</u> 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*)

<u>Level of Analysis: Program, 4 Measured Entities</u>

1. Texas Medicaid
   A. 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. Florida CHIP
   A. 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)

3. Commercial
    A. 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

4. Florida Medicaid
    A. Size: # Members 0-20 years, CY 2010: 2,068,670
    B. 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.

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

2) FL CHIP – Plan 2
    A. 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 three 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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2011				
	Program 1	Program 2	Program 3	Plan 1	Plan 2
Total Number Patients	3,544,247	317,146	184,152	140,986	168,191
Age Group Distribution					
Age <1 years	7.05%	N/A	1.54%	N/A	N/A
Age 1-2 years	14.32%	N/A	5.75%	N/A	N/A
Age 3-5 years	19.46%	3.81%	12.68%	4.12%	3.60%
Age 6-7 years	11.21%	13.05%	9.57%	13.71%	12.55%
Age 8-9 years	9.85%	15.00%	10.18%	15.76%	14.41%
Age 10-11 years	9.03%	15.71%	10.55%	16.27%	15.25%
Age 12-14 years	11.63%	23.73%	16.09%	23.06%	24.31%
Age 15-18 years	13.19%	28.70%	22.13%	27.08%	29.88%
Age 19-20 years	4.27%	N/A	11.50%	N/A	N/A
Geographic Location					
Urban	83.63%	92.94%	95.95%	93.01%	92.91%
Rural	15.15%	5.02%	3.86%	4.83%	5.15%
Missing	1.22%	2.04%	0.19%	2.16%	1.94%
Race and Ethnicity					
Non-Hispanic White	17.36%	N/A	N/A	N/A	N/A
Non-Hispanic Black	15.08%	N/A	N/A	N/A	N/A
Hispanic	58.07%	N/A	N/A	N/A	N/A
Other & Unknown	9.49%	N/A	N/A	N/A	N/A

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

	Descriptive Characteristics of Individuals 0-20 Years Enrolled at Least One Month, CY 2010					
	Program 1	Program 2	Program 3	Program 4	Plan 1	Plan 2
Total Number Patients	3,393,963	315,975	189,968	2,068,670	77,255	116,388
Age Group Distribution						
Age <1 years	7.35%	N/A	1.45%	6.05%	N/A	N/A
Age 1-2 years	15.16%	N/A	5.67%	14.23%	N/A	N/A
Age 3-5 years	19.48%	3.64%	12.73%	19.26%	5.72%	4.22%
Age 6-7 years	11.12%	13.32%	9.69%	10.47%	15.68%	12.54%
Age 8-9 years	9.70%	15.14%	10.24%	9.19%	16.99%	14.21%
Age 10-11 years	8.75%	15.84%	10.60%	8.74%	16.41%	15.18%
Age 12-14 years	11.23%	23.70%	16.20%	11.87%	21.40%	24.05%
Age 15-18 years	12.99%	28.37%	22.12%	14.73%	23.79%	29.81%
Age 19-20 years	4.22%	N/A	11.31%	5.47%	N/A	N/A
Geographic Location						
Urban	83.20%	92.08%	96.70%	91.47%	92.10%	92.11%
Rural	15.56%	5.07%	3.17%	7.30%	5.00%	5.19%
Missing	1.24%	2.85%	0.13%	1.23%	2.89%	2.70%
Race and Ethnicity						
Non-Hispanic White	18.21%	N/A	N/A	29.89%	N/A	N/A
Non-Hispanic Black	15.45%	N/A	N/A	29.39%	N/A	N/A
Hispanic	59.42%	N/A	N/A	29.65%	N/A	N/A
Other & Unknown	6.92%	N/A	N/A	11.06%	N/A	N/A

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 program. Texas has the third largest Medicaid program in the U.S. 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 program 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) the evidence regarding the effectiveness of professionally applied topical fluoride in caries prevention, (2) an environmental scan, 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. Other Threats to Reliability - Sample Size

Our measured entities include very large numbers of patients; 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

Topical Fluoride measures the percentage of children aged 1-21 years at moderate to high risk for dental caries who had at least 2 topical fluoride applications during the reporting year. 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) 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 program. 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 & Methodology

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

# Requested	# Received	%Received
600	414	69%

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 topical fluoride application 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 – Topical Fluoride Application Procedure Codes

Data Extraction. For the second component of validation, assessing whether the specific preventive service of topical fluoride application is accurately captured by claims data, chart abstraction forms were developed by the research team. 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 fluoride application 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 topical fluoride procedure codes. 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 - IDENTIFYING ELEVATED RISK WITH CLAIMS DATA

Evidence based guideline indicate that fluoride is most effective for children at higher risk for caries. Thus, inclusion in the denominator is limited to children identified as being at moderate to high risk for caries. Administrative claims data for dental claims typically do not include diagnostic codes. Procedure codes for risk assessment that identify moderate and high risk were included in the measure logic. However, because these are newer codes, additional logic was included to identify children with recent history of restorations, which are indicative of caries. A systematic review found that prior caries experience to be an important predictor of future risk (Zero D, Fontana M, Lennon AM. 2001. Clinical applications and outcomes of using indicators of risk in caries management. J Dent Educ. 2001 Oct;65(10):1126-32.) Expert consensus and validation through chart reviews was done to finalize the procedure codes (indicated in the measure specifications) used to identify elevated risk. The test data results reported in this application demonstrate that it is feasible to use these validated codes to identify children at elevated risk who should receive preventive services.

4. 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

Number of Procedure Codes	Record and Procedure Code on Claim Correlate	Record Did Not Correlate with Procedure Code on Claim	Unable to Determine Correlation
1,135	94.04%	4.22%	1.75%

B. Critical Data Element Validation – Topical Fluoride Application Procedure Codes

To assess whether the specific preventive service of topical fluoride application is 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 for topical fluoride applications. Agreement (concordance) for topical fluoride application was 89.9%. Sensitivity was 90.7% and specificity was 88.4%. The positive predictive value was 93.5% and negative predictive value was 83.9%. 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 0.782, which is at the high end of the "substantial agreement" category.

Table 2b2.3-2 Agreement between Record and Administrative Data for Specific Services

	Concordance	Prevalence	Sensitivity	Specificity	PPV	NPV	Kappa
Fluoride	89.91%	0.647	0.907	0.884	0.935	0.839	0.782
Dates of service: 317			(0.857-0.942)	(0.806-0.934)	(0.888-0.963)	(0.757-0.898)	(0.710-0.853)
# indeterminate: 0							

95% confidence intervals indicated in parentheses

Our findings are similar 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. For fluoride, they found lower sensitivity (80%), higher specificity (98%) and similar kappa value (0.81) of billing data compared to direct observation. (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

The measures concept of preventive dental services identified using CDT codes (within which topical fluoride falls) was identified through the Delphi rating process as a high-scoring measure concept with a mean importance score of 7, mean feasibility score of 8, and mean validity score of 7 for specific evidence-based preventive services, 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.] Thus, the measure has face validity. However, gaps were identified with existing preventive services measures, including defining "preventive services" too broadly (encompassing services without sound evidence of their effectiveness in caries prevention), lack of clear specifications and lack of standardization. Although the scan included two measure concepts that were specific to fluoride, they were deemed to be low scoring because they pertained to "fluoride supplements" or "fluoride exposure assessment." Scientific soundness was limited due to lack of clarity in measure description.

Content Validity. In addition, the measure also demonstrates **content validity** – the extent to which the measure specifications reflect the intended domain of care. This measure directly reflects evidence-based guidelines regarding an effective caries prevention measure (professionally applied topical fluoride), including the frequency required for clinical effectiveness (at least every three-six months). Please see the Measure Evidence Form for more details.

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?*)

As noted above, the overall agreement between the administrative claims data and dental record data was high based on both simple agreement and using the more conservative Kappa statistic. Overall, we interpret these findings as evidence that validates the accuracy of administrative claims data for performance measurement purposes. These empirical findings, combined with our face validity and content validity assessments of the measure score, lead us to conclude that both the data elements and the measure score represent valid measures of the evidence-based preventive service topical fluoride application.

2b3. EXCLUSIONS ANALYSIS
NA X☐ no exclusions — *skip to section 2b4*

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.
Not applicable.

2b4.1. What method of controlling for differences in case mix is used?
☐X No risk adjustment or stratification
☐ Statistical risk model with _risk factors
☐ Stratification by _risk categories
☐ Other,

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*)
Not applicable

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

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 four 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 3 programs with performance data for 2011, (2) the 4 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 two topical fluoride applications, the dichotomous outcome of had/did not have two topical fluoride applications 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/Plan, Year, Measure Score as % (Measure Score, SD, Lower 95% CI, Upper 95% CI)

Program 1, CY 2011:	37.13%	(	0.3713 ,	0.0004 ,	0.3704 ,	0.3722)
Program 2, CY 2011:	27.15%	(	0.2715 ,	0.0020 ,	0.2676 ,	0.2754)
Program 3, CY 2011:	22.04%	(	0.2204 ,	0.0020 ,	0.2165 ,	0.2243)
Program 1, CY 2010:	34.96%	(	0.3496 ,	0.0005 ,	0.3487 ,	0.3505)
Program 2, CY 2010:	22.63%	(	0.2263 ,	0.0019 ,	0.2225 ,	0.2301)
Program 3, CY 2010:	35.04%	(	0.3504 ,	0.0023 ,	0.3458 ,	0.3550)
Program 4, CY 2010:	18.16%	(	0.1816 ,	0.0009 ,	0.1799 ,	0.1833)
Plan 1, CY 2011:	25.50%	(	0.2550 ,	0.0030 ,	0.2491 ,	0.2609)
Plan 2, CY 2011:	28.69%	(	0.2869 ,	0.0027 ,	0.2815 ,	0.2923)
Plan 1, CY 2010:	23.24%	(	0.2324 ,	0.0048 ,	0.2230 ,	0.2418)
Plan 2, CY 2010 :	23.76%	(	0.2376 ,	0.0034 ,	0.2309 ,	0.2443)

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 in both years and between plans in one of the two years (Table 2b5.2).

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

	Chi-Square Value	*p* - value
Program Results, 2011	5887.1	<0.0001
Program Results, 2010	23554.5	<0.0001
Plan Results, 2011	61.2	<0.0001
Plan Results, 2010	0.8	0.3711

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, with program-level performance scores ranging by approximately 17 percentage points. At the plan level, statistically significant differences were detected in 2011, but not in 2010. This is consistent with a greater difference in performance between the two plans in 2011 (25.50% and 28.69%) than in 2010 when the rates were almost equal (23.24% and 23.76%). This is precisely the purpose of performance measurement - to detect when there are differences in performance. In 2010, there was no appreciable difference in performance between the two plans. Collectively, however, it is clear that this measure detects differences in performance on the measure scores when they do exist. Our findings are consistent with evidence reported elsewhere in this application documenting a performance gap and disparities in performance. Thus, Topical Fluoride informs performance improvement efforts by allowing plans and programs to identify and monitor performance gaps and disparities 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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
This measure is specified for reporting at the program and plan level and there are currently no efforts to develop an eMeasure (eCQM) of this measure at these levels.

Our understanding is that the Feasibility Score Card is only for eMeasures; consequently, we have not submitted this. Feasibility criteria were met during the initial endorsement review.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
This measure relies on standard data elements in administrative claims data (e.g., patient ID, patient birthdate, enrollment information, CDT codes, date of service, and provider taxonomy). These data are readily available and can be easily retrieved because they are routinely used for billing and reporting purposes. A key advantage of using administrative claims data is that the time and cost of data collection for performance measurement purposes are relatively low because these data are already collected for other purposes.

Initial feasibility assessments were conducted using the RAND-UCLA modified Delphi process to rate the measure concepts with feasibility as one component of the assessment. On a 1-9 point scale, the measure concept of preventive dental services identified using CDT codes (within which topical fluoride falls) was rated as an 8 or "definitely feasible" by the expert panel. During the empirical testing phase, our testing found that the critical data elements had missing/invalid data of <1% (Data 3c.1.), meeting or exceeding the guidance from the Centers for Medicare and Medicaid Services regarding acceptable error rates. During measure development and testing, the measure specifications were made available through a publicly accessible website for public comment with additional broad email dissemination to a wide range of stakeholders. No concerns regarding feasibility were raised during this process.

Citation: Centers for Medicare & Medicaid Services. Medicaid and CHIP Statistical Information System (MSIS) File Specifications and Data Dictionary. 2010; http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MSIS/downloads/msisdd2010.pdf. Accessed August 10, 2013.

Data 3c.1 Percentage of Missing and Invalid Values for Critical Data Elements

PROGRAM 1
Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.01%
Rendering Provider ID: 0.28%

PROGRAM 2
Member ID: 0.27%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.28%
Date of Service: 0.00%
Rendering Provider ID: 0.18%

PROGRAM 3

Member ID: 0.00%
Date of Birth: 0.00%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.01%
Date of Service: 0.00%
Rendering Provider ID: 0.61%

PROGRAM 4
Member ID: 0.43%
Date of Birth: 0.02%
Monthly enrollment indicator: 0.00%
Dental Procedure Codes - CDT: 0.00%
Date of Service: 0.00%
Rendering Provider ID: 0.67%

Endorsement Maintenance Update: There have been no reports of feasibility issues with implementing this measure. Please see Use and Usability section.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
This measure is intended to be transparent and available for widespread adoption. As such, it was purposefully designed to avoid using software or other proprietary materials that would require licensing fees. The measure specifications, including a companion User Guide, is accessible through a website and available free of charge for non-commercial purposes. The main requirements of users is to ensure the quality of their source data and expertise to program the measures within their information systems, following the clear and detailed specifications. Technical assistance is available to users.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting Texas Health and Human Services Commission: Texas Medicaid and CHIP Texas Health and Human Services Commission: Texas Medicaid and CHIP Payment Program https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf Texas Health and Human Services Commission: Texas Medicaid and CHIP Quality Improvement (external benchmarking to organizations) Covered California

	http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf Quality Improvement (Internal to the specific organization) State Medicaid Agencies http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/? Michigan Healthy Kids Dental RFP https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids

4a1.1 For each CURRENT use, checked above (update for <u>maintenance of endorsement</u>), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

1. Program and Sponsor: Texas Health and Human Services Commission - Texas Medicaid and CHIP

https://hhs.texas.gov/sites/default/files//documents/laws-regulations/handbooks/umcm/6-2-15.pdf

Purpose: Payment Program and Public Reporting

This measure has been adopted by the Texas Health and Human Services Commission as part of the Texas CHIP and Medicaid Dental Services Pay-for-Quality (P4Q) program. [Texas HHSC Uniform Managed Care Manual, Chapters 6.2.15. Effective Date 09/01/2017, Version 2.0].

This measure was also present in earlier iterations of the Texas Medicaid and CHIP quality programs since initial endorsement. We are referencing current use for this update.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies to the state of Texas CHIP and Medicaid programs (statewide application). There are two dental plans (i.e., the accountable entities) that serve Texas CHIP and Medicaid. In June 2017, there were 3,359,770 children enrolled in Texas Medicaid and CHIP (https://hhs.texas.gov/about-hhs/records-statistics/data-statistics/healthcare-statistics).

Level of Measurement and Setting: The measure is implemented at the plan and program level within the Texas Medicaid and CHIP programs.

2. Covered California, the California Health Benefit Exchange

http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-Individual-Model-Contract.pdf
http://hbex.coveredca.com/insurance-companies/PDFs/2017-2019-QDP-Issuer-Contract-and-Attachments.pdf

Purpose: Quality Improvement

This measure is included in the Covered California Qualified Health Plan Issuer Contract for 2017-019 For the Individual Market and the Covered California Qualified Dental Plan Issuer Contract for 2017-2019. The measure is to be reported annually.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
This applies statewide. In March 2017 there were 85,000 enrollees 0-18 years old in CC health plans (which may offer dental benefits and would therefore report on the dental quality measures). There were 5,100 children enrolled specifically in Qualified Dental Plans. (http://hbex.coveredca.com/data-research/)

Level of Measurement and Setting. The measure is implemented at the plan level with the Covered California program.

3. State Medicaid Agencies

http://www.msdanationalprofile.com/2015-profile/management-reporting-and-quality-measurement/quality-measurement/?

(Note: To access the data, a public user account must be created. We can help facilitate access to the data if needed.)

Purpose: Quality Improvement

The Medicaid|Medicare|CHIP Services Dental Association conducts an annual survey of state Medicaid programs and collects data specifically on which programs report Dental Quality Alliance measures.

In its 2015 profile (the most recent available), 9 states reported that they currently use this measure in the Medicaid and/or CHIP programs.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The 9 states are: Alabama, Connecticut, Florida, Idaho, Illinois, Nevada, Oklahoma, Rhode Island, and West Virginia. Data are not provided on the number of accountable entities included.

4. Michigan Healthy Kids Dental Program
https://www.buy4michigan.com/bso/external/bidDetail.sdo?bidId=007117B0011386&parentUrl=activeBids

Note: Select Schedule A Work Statement link under File Attachments

Purpose: Quality Improvement

The Michigan Healthy Kids Dental Program has included this measure in the set of measures included in its Performance Monitoring Standards, which is currently included in the Request for Proposals and will be included in the contracts between the contracted dental plans and the State of Michigan.

Geographic Area and Number/Percentage of Accountable Entities and Patients:
The Healthy Kids dental program covers children enrolled in Michigan's Medicaid program statewide. The state intends to award two contracts. There are approximately 955,000 enrollees served by the Healthy Kids Dental Program.

Additional Information:
This measure was one of ten performance measures that focused on Dental Caries Prevention and Disease Management among children and that was approved by the DQA. 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)

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not applicable.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Per the annual survey conducted by the Medicaid|Medicare|CHIP Services Dental Association (MSDA), 9 Medicaid/CHIP agencies are implementing this measure. The measure is part of measure set included in the Request for Proposal (RFP) released by the Michigan Healthy Kids Dental Program. This measure is included in the Pay-For-Quality program and publicly reported in the Texas Medicaid and CHIP programs. Additionally, this measure is a requirement for the Qualified Dental Plans to report to the Covered California, the state-based marketplace in California.

The DQA provides technical assistance to these and other users of DQA measures through webinars, resource document development, and one-on-one staff support. The DQA has an Implementation Committee dedicated to developing implementation and improvement resources.

In order to ensure transparency, incorporate learnings from implementation, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures are just getting underway and there is limited data reporting. Implementation has mostly focused on addressing questions related to how to use the measures in the context of broader quality improvement and clarifying questions related to the specifications.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
In an effort to facilitate implementation of the DQA measures, the DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
In order to ensure transparency, establish proper protocols for timely assessment of the evidence and measure properties, and to comply with the NQF's endorsement agreement, the DQA has established an annual measure review and maintenance process. This measure review process is overseen by the DQA's Measures Development and Maintenance Committee (MDMC) which is comprised of subject matter experts. This annual review process includes: (1) call for public comments, (2) evaluation of the comments, (3) user group feedback, and (4) code set reviews.

DQA provides technical assistance on an ongoing basis to users of DQA measures through webinars, resource document development and one-on-one staff support.

In 2016, the DQA expanded its scope of review of its measures by convening conference calls for two user groups – one comprised of representatives from 6 state Medicaid programs (Alabama, Florida, Kentucky, Oregon, Nevada, and Pennsylvania) and the other comprised of representatives from 8 dental plans. Participants shared their experiences implementing DQA measures in their respective programs, including any challenges related to the DQA measures specifications and use of these measures in their quality improvement programs. Participants did not have any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
There has been no feedback indicating any significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.2.3. **Summarize the feedback obtained from other users**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
There have been no significant issues related to the clarity or feasibility of implementing the measure specifications.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
This is the first 3-year maintenance endorsement review for this measure. As indicated above, the measure is being implemented in multiple programs. Because measure implementation requires a start-up phase for integration of the measures into contracts and for programs and plans to prepare for reporting, in combination with a lag period for reporting measures calculated using administrative claims data, most of the entities that have adopted the measures either have only limited baseline scores or will start reporting measures within the next year.

We are only aware of repeat measurements within the Texas Medicaid/CHIP programs (https://thlcportal.com/qoc/dental), which started implementing this measure after it was approved by the Dental Quality Alliance and before NQF endorsement, as follows:

Texas Medicaid
Year, Program Denominator, Program Overall Score, DentaQuest(Plan) Score, MCNA(Plan) Score
2014, 1090952, 39.97, 41.57, 37.62
2015, 1334887, 41.75, 44.70, 38.15

Texas CHIP
Year, Program Overall, DentaQuest(Plan), MCNA(Plan)
2014, 108704, 33.01, 35.45, 32.99
2015, 79693, 37.50, 41.44, 37.71

These data suggest a trend in improvement over time. However, as noted above, these are initial performance data for one program. Most measure users are just now getting their quality measurement programs underway.

4b2. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
No unintended or negative consequences have been identified.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization of Related Measures
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable.

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** Appendix_Fluoride.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Dental Association on behalf of the Dental Quality Alliance
Co.2 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-
Co.3 **Measure Developer if different from Measure Steward:** American Dental Association on behalf of the Dental Quality Alliance
Co.4 **Point of Contact:** Krishna, Aravamudhan, aravamudhank@ada.org, 312-440-2772-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This project is headed by the DQA through its Measure Development and Maintenance Committee (formerly Research and Development Committee). The following individuals were responsible for executing and overseeing all scientific aspects of this project.

- Craig W. Amundson, DDS, General Dentist, HealthPartners, National Association of Dental Plans. Dr. Amundson serves as chair for the Committee.
- Mark Casey, DDS, MPH, Dental Director, North Carolina Department of Health and Human Services Division of Medical Assistance
- Natalia Chalmers, DDS, PhD, Diplomate, American Board of Pediatric Dentistry, Director, Analytics and Publication, DentaQuest Institute
- Frederick Eichmiller, DDS, Vice President & Science Officer, Delta Dental of Wisconsin
- Chris Farrell, RDH, BSDH, MPA, Oral Health Program Director, Michigan Department of Health and Human Services

This group oversees the maintenance. All work of this Committee was distributed for review and formal vote and approval by the entire Dental Quality Alliance. (http://ada.org/dqa) The DQA is made up of representatives from 38 stakeholder organizations.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2013
Ad.3 **Month and Year of most recent revision:** 01, 2017
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2018

Ad.7 **Disclaimers:** Dental Quality Alliance measures and related data specifications, developed by the Dental Quality Alliance (DQA), are intended to facilitate quality improvement activities. These Measures are intended to assist stakeholders in enhancing quality of care. These performance Measures are not clinical guidelines and do not establish a standard of care. The DQA has not tested its Measures for all potential applications.
Measures are subject to review and may be revised or rescinded at any time by the DQA. The Measures may not be altered without the prior written approval of the DQA. The DQA shall be acknowledged as the measure steward in any and all references to the measure.
Measures developed by the DQA, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and DQA. Neither the DQA nor its members shall be responsible for any use of these Measures.
THE MEASURES ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND
Limited proprietary coding is contained in the Measure specifications for convenience.
For Proprietary Codes:

This material contains National Uniform Claim Committee (NUCC) Health Care Provider Taxonomy codes

(http://www.nucc.org/index.php?option=com_content&view=article&id=14&Itemid=125).
Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. The DQA, American Dental Association (ADA), and its members disclaim all liability for use or accuracy of any terminologies or other coding contained in the specifications.
THE SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** In 2008, the Centers for Medicare and Medicaid Services (CMS) asked the ADA to lead the development of a broad coalition of organizations that would lead dentistry to improve the oral health of Americans through quality measurement and quality improvement. The ADA subsequently established the DQA. The DQA is a multi-stakeholder alliance comprised of approximately 38 stakeholders (with organizations as members) from across the oral health community, including federal agencies, third-party payers, professional associations, and an individual member from the general public. The DQA's mission is to advance the field of performance measurement to improve oral health, patient care, and safety through a consensus building process.