NATIONAL QUALITY FORUM

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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

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

Staff Preliminary Analysis: Maintenance of Endorsement

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)

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.]

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

1. Evidence and Performance Gap – 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

1a 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)"

<table>
<tr>
<td>Grade assigned to the evidence associated with the recommendation with the definition of the grade</td>
<td>
"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
<table>
<tr><th>Study Design</th><th>Adapted Quality Criteria</th></tr>
<tr><td>Randomized and non-randomized controlled-trials, adapted from the U.S Preventive Services Task Force methods (Harris et al, 2001)</td><td>• 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?</td></tr>
</table>
• 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.
</td>
</tr>
</table>

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.

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

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 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*)

Sample for measure score reliability testing: 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.

Sample for construct validity testing: 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 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*)

Patient sample for measure score reliability testing: 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 <u>performance measure score</u> 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)*

<u>Method of testing construct validity</u>: 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.

<u>Method of assessing face validity</u>: 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
Commercial	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
Medicaid	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 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

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.6 Copyright statement: © by the National Committee for Quality Assurance

1100 13th Street, NW, 3rd Floor

Washington, DC 20005

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 noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA. © 2017 by the National Committee for Quality Assurance