NATIONAL QUALITY FORUM

Memo

TO: Eye Care, Ear, Nose and Throat (EENT) Standing Committee

FR: NQF Staff

RE: Post-Comment Call

DA: June 5, 2017

Purpose of the Call

The Eye Care, Ear, Nose and Throat (EENT) Standing Committee will meet via conference call on Tuesday, June 13, 2017 from 1:00 pm to 3:00 pm ET. The purpose of this call is to:

- Re-vote on Scientific Acceptability criteria for measure 2640: Otitis Media with Effusion – Antibiotics Avoidance and potentially vote on other criteria

Standing Committee Actions

1. Review this briefing memo and Draft Report
2. Be prepared to re-vote on Reliability sub-criteria for measure #2640 and to consider voting for Validity sub-criteria. The complete measure submission worksheet is provided in Appendix A and additional testing information provided after the submission deadline in provided in Appendix B. The evaluation summary of the measure from the draft report are provided in Appendix C.

Conference Call Information

Please use the following information to access the conference call line and webinar:
Speaker dial-in #: (855) 696-3824 (**Committee only**. No conference code required.*)*
Public dial-in #: (877) 315-9042
Web Link: http://nqf.commpartners.com/se/Rd/Mt.aspx?309566

***In order to vote, Committee members must use their individual webinar links sent via email.**

Background

The EENT Standing Committee's spring 2017 off-cycle activity included evaluating two newly-submitted measures against NQF's standard evaluation criteria. The Committee recommended one measure for endorsement:

- 2811: Acute Otitis Media – Appropriate First-Line Antibiotics

The Committee did not recommend the following measure:

- 2640: Otitis Media with Effusion - Antibiotics Avoidance

Comments Received

NQF solicits comments on measures undergoing review in various ways and at various times throughout the evaluation process. First, NQF solicits comments on endorsed measures on an ongoing basis through the Quality Positioning System (QPS). Second, NQF solicits member and public comments prior to the evaluation of the measures via an online tool located on the project webpage. Third, NQF opens a 30-day comment period to both members and the public after measures have been evaluated by the full Committee and once a report of the proceedings has been drafted.

Pre-evaluation comments

The pre-evaluation comment period was open from February 24 to March 10, 2017 for the measures under review. Two pre-evaluation comments were received: 1.) harmonization was encouraged between NQF #0657 and newly-submitted #2640 before further consideration of endorsement and 2.) AAO-HNSF highlighted the difference between the American Academy of Pediatrics Clinical Practice Guideline, Diagnosis and Management of Acute Otitis Media, and the denominator of measure #2811. All pre-evaluation comments were provided to the Committee prior to their initial deliberations.

Post-evaluation comments

The Draft Report was released for Public and Member comment from April 27 to May 30, 2017. During this commenting period, NQF received no additional comments. We have included all of the comments that we received in the Comment Table. This table contains the commenter's name, comment, and associated measure.

Committee Re-Vote on Measure #2640

#2640: Otitis Media with Effusion - Antibiotics Avoidance

During the evaluation of this measure, the Standing Committee voted against endorsement, primarily due to concerns with the difficulty of diagnosing otitis media with effusion and the potential for providers to game the measure. After a lengthy discussion, the Committee agreed that the measure did not pass the reliability subcriterion.

In considering the discussion, NQF staff determined that the committee's discussion regarding ability to diagnosis OME is more properly one of validity rather than reliability. NQF will ask the Committee to re-vote on reliability, basing its rating on clarity of specifications and results of reliability testing, and to consider the question of diagnosis accuracy and the data provided by the developer in a discussion on validity. If the measure passes both the reliability and validity subcriteria, members will then re-discuss and re-vote on subcriterion 1b (opportunity for improvement) and then discuss and vote on the remaining criteria.

Appendix A: Measure Worksheet - #2640

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 2640
Measure Title: Otitis Media with Effusion - Antibiotics Avoidance
Measure Steward: The Children´s Hospital of Philadelphia Pediatric Quality Measures Program Center of Excellence
Brief Description of Measure: The proportion of encounters with a diagnosis of Otitis Media with Effusion (OME) made at age 2 months to 12 years, where patients were not prescribed systemic antimicrobials.
Developer Rationale: Otitis media is a highly prevalent condition among young children particularly, and a major driver of outpatient health care and antibiotic utilization. Otitis Media with Effusion (OME), a non-infectious chronic condition, occurs in 6.5% of children 0-11 years of age in our test population; it is well-established that, unlike acute otitis media, antibiotic therapy provides no benefit for OME. Nonetheless, antibiotics have been frequently prescribed in this context. Reduction of inappropriate antibiotic use has significant public health benefits, including reduced development of antibiotic resistance, reduced health care cost, and decreased side effect burden for patients.

This measure is developed for evaluation in electronic health record data, providing quality measurement across a large range of population sizes and data types. The existing Academy of Otolaryngology-Head and Neck Surgery (AAOHNS)/Academy of Family Physician´s (AAFP) PCPI measures are specified as manual chart review measures, limiting sample size for evaluation. They also used uncommon coding systems, which was the proximate cause of the failure of the initial CHIPRA core set OME measure. We are addressing both issues, and testing with data from multiple EHR systems.

Numerator Statement: Eligible encounters at which a systemic antibiotic was not prescribed.
Denominator Statement: Outpatient encounters at which otitis media with effusion is diagnosed, but at which common conditions for which antibiotics are indicated are not diagnosed. It is expected that a small fraction of patients with rare non-OME indications for antibiotic usage will not be identified by the specified exclusion criteria, but these will be rare cases, and will not alter the measure score significantly in most practice contexts. Of note, however, applicability may be limited in specific practice environments in which a large proportion of patients seen have immune deficiencies requiring chronic antibiotic use (e.g. immunology or hematology/oncology clinics).
Denominator Exclusions: Diagnosis at the visit of common childhood infection for which antibiotics are frequently indicated.

Measure Type: Process
Data Source: Other
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Facility, Integrated Delivery System

New Measure -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality.

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.

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

The developer provides the following evidence for this measure:

- The evidence for this process measure is based on a clinical practice guideline recommendation from two peer reviewed publications: The Pediatrics Journal (2004) and Otolaryngology Head and Neck Surgery Journal.
- This is a strong recommendation.
  - The Otolaryngology article states: "ANTIBIOTICS: Clinicians should recommend against using systemic antibiotics for treating OME. Strong recommendation against based on systematic review of RCTs and preponderance of harm over benefit."
- The systematic review for the clinical practice guideline assessed the Quality, Quantity and Consistency (QCC) of literature based on 20 systematic reviews and 49 randomized controlled trials (RCTs). The aggregate evidence quality was assessed as ***Grade A***.
- The developer estimates the following benefits over harm in the recommendation against therapy:
  - avoidance of side effects and reduction in cost by not administering medications;
  - avoidance in delays in definitive therapy caused by short-term improvement then relapse; and
  - avoidance of societal impact of inappropriate antibiotic prescribing on bacterial resistance and transmission of resistant pathogens.

Exception to evidence

N/A

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm

Process measure based on systematic review and grading of evidence (Box 3) → QQC presented (Box 4) → Quantity: high; Quality: High; Consistency: High (5a) → High

The highest rating possible is HIGH.

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

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

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

The developer provides the following information:

- This measure was tested on 36,060 encounter records from CHOP EHR from 2009 to 2014.
  - Score characteristics for provider-specific evaluation (N=531; with 285 representing providers with 5 or more visits) were as follows: mean failure rate 15.05%; median 8.00%; IQR 3.00% - 20.00%.
  - Score characteristics for specialty-specific evaluation (N=3) were as follows: mean failure rate 11.42%; median 6.00%; IQR 0.00% - 16.00%.

Disparities

- The developers reported finding relatively small, but statistically significant, differences in provider-level performance between racial/ethnic groups and those with varying insurance status/type. However, they did not provide the data from these analyses.

Questions for the Committee:

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

Committee pre-evaluation comments

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

1a

- The evidence for this process measure is based on a clinical practice guideline recommendation from two peer reviewed publications: The Pediatrics Journal (2004) and Otolaryngology Head and Neck Surgery Journal.
- This is a strong recommendation.
- The systematic review for the clinical practice guideline assessed the Quality, Quantity and Consistency (QCC) of literature based on 20 systematic reviews and 49 randomized controlled trials (RCTs). The aggregate evidence quality was assessed as Grade A.
- What is the relationship of this measure to patient outcomes? Direct overuse harms pts
- How strong is the evidence for this relationship? Robust
- Is the evidence directly applicable to the process of care being measured? Yes
- The process being measured is withholding antibiotics to pediatric patients with OME where the numerator is the number of encounters in which an antibiotic was not prescribed and denominator is the diagnosis of common childhood infection for which antibiotics are frequency indicated. The measure is tangential as it is not specific to just OME.
- With overuse measures should the numerator be number of patient who were prescribed abx? Also, how do the practitioners differentiate AOM and OME? What are the guidelines? Does "common childhood infections" need to be better defined. Do children with immune deficiency or chronic disorders need to be better defined?
- OME occurs in 6.5% of children and antibiotics are not recommended. Evidence-clinical practice guideline.
- The evidence of the lack of effect of abx in OME directly relates to the process of care being measured. It applies directly. The process being measured relates directly to the desired outcome (of not prescribing abx for OME)

1b.

- Don't fully understand CHOP analysis but does seem like room for improvement. Disparities appear minor.
- There was a large difference between the mean and median but both numbers were small. Is the gap significant?
- Gaps in care were not addressed.
- The performance gap is the weakest link in the chain here. There is some newer data that the performance gap might not be large enough to warrant the measure. There are small disparities that are not large enough to affect the measure.

1c.

- high prevalence and significant issue. If this could reduce abx would be very + outcome.
- Over prescription of antibiotics resulting in high cost, resistance to antibiotics, and side effect burdens.
- reduction in abx resistance, cost savings, and decreased treatment burden.
- The problem is antibiotic resistance is the high severity problem and high cost problem addressed by the measure.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Computable eMeasure; EHR does not implement HQMF

Specifications:

- This measure is specified for the individual and group clinician, facility, and integrated delivery system levels of analysis, for use in the clinician office/clinic and urgent care settings. A higher score indicates better quality.
- The numerator of this measure is eligible encounters at which a systemic antibiotic was not prescribed.
- The denominator of this measure is outpatient encounters at which otitis media with effusion is diagnosed, but at which common conditions for which antibiotics are indicated are not diagnosed. Patients included in the denominator are those ages 2-155 months, inclusive, at the date of the encounter.
- The denominator exclusions include diagnosis at the visit of common childhood infection for which antibiotics are frequently indicated.
- Applicable diagnoses used in the measure denominator and exclusions are identified using ICD-9-CM, ICD-10-CM, and SNOMED-CT codes. Antibiotics used in the measure numerator are identified using RxNorm codes. ***All codes are listed in an Excel file (OME_VSAC_ValueSets.xls) included with the submission materials***.
- To be eligible for the measure, a provider must have more than 5 eligible encounters in the measurement time period. The time period itself (e.g., calendar year, quarterly, etc.) is not specified.
- The calculation algorithm is stated in S.18 and appears straightforward.

Questions for the Committee:

- *Is it likely this measure can be consistently implemented?*
- *Are there any concerns with not specifying a measurement timeframe?*

eMeasure Technical Advisor(s) review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	Submitted eMeasure contains components that cannot be represented due to limitations of HQMF or QDM and the developer explained the work around for these limitations. It appears that the developer did not use the Measure Authoring Tool to generate the XLM code and HQMF format. However, the metatags in the format submitted are identical to HQMF and they align to the QDM.

Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC.
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously.
Feasibility Testing	The submission contains a feasibility assessment that addresses data element feasibility and follow-up with measure developer indicates that the measure logic is feasible based on assessment by EHR vendors**.**

2a2. Reliability Testing, Testing attachment

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.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data used in testing was obtained from 6 academic pediatric health systems, which used 3 different EHR vendors (Epic Systems EHR, Cerner's Millennium EHR, and a combination of Allscripts' EHR and an institutional Emergency Department system).
    - These data included records from January 2009-June 2016 and included information for 704 clinicians and 207 practices.
    - All patients with at least one evaluable visit were included in testing.

Site	N	Male %	White %	Black %	Other race %	0-2 y %	2-5 y %	6-12 y %
A	19,070	57.49	61.63	23.29	15.08	41.86	40.77	17.34
B	30,713	58.33	62.41	17.14	20.45	31.07	47.73	21.20
C	11,726	58.18	65.82	5.09	29.09	36.98	43.46	19.56
D	29,258	58.35	69.42	12.71	17.87	37.12	42.04	20.83
E	6,079	59.16	77.40	15.92	6.68	38.95	43.66	17.39
F	7,803	58.10	59.87	6.33	33.80	34.74	41.59	23.66

- The developer conducted score-level reliability testing by conducting an Analysis of Variance (ANOVA) to test whether measure results were statistically significantly different. This is an appropriate method of testing score-level reliability. Note that because clinicians and practices had a different number of visits, the value 1-1/F can be considered an "average reliability". A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 is often regarded as a minimum acceptable reliability value.
    - NOTE that NQF requires testing for all levels of analysis for which the measure is specified (in this case, for clinicians, practices, facilities, and systems). Although the ***overall method*** (ANOVA) is appropriate, the developers appear to have aggregated data at two levels to conduct the ANOVA (e.g., for providers, then by site), and therefore the results do not seem to demonstrate differences for the four levels of analysis. Moreover, additional testing would be needed before conferring endorsement for all four levels of analysis, as only two sets of results were provided. NQF staff have requested additional information from the developer, but it is not yet available.
- Developers provided pictorial representation of the variations between clinicians and practices and over time.
    - These graphics suggest that a few clinicians and practices—those with very low measure scores—likely can be distinguished from other providers, but they do not indicate whether other clinicians and practices can be differentiated.

- The developer did not conduct data element reliability testing. NQF agrees with the developer that data element reliability testing that assesses consistency of calculation is not needed for an eMeasure, which by definition should be calculated consistently.

Results of reliability testing

- Because the ANOVAs appear to be conducted on aggregated data rather than for each level of analysis as specified, additional information from the developer is required before the results can be interpreted.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- ***If** the aggregation for the ANOVA is appropriate, do the results demonstrate sufficient reliability so that conclusions about quality can be made?*

Guidance from the Reliability Algorithm

Specifications are precise (Box 1) → Empirical testing conducted, although testing at additional levels of analysis will be needed (Box 2) → Score-level testing was conducted for 2 of the 4 specified levels of analysis (Box 4) → Method does not appear to be appropriate (Box 5) → Insufficient or Low

The highest possible rating is HIGH.

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

RATIONALE: It appears that the method is not quite appropriate, but additional information from the developer is needed. Also, additional testing will be needed in order to endorse the measure for all four levels of analysis specified. If the method is not appropriate, the rating for the criteria would be LOW, unless the developer can provide additional analysis.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Do you agree that the specifications are consistent with the evidence?*

2b2. Validity testing

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.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity**
☒ **Empirical validity testing of the measure score**

Validity testing method:

- Score-level testing

- Developers compared measure results across the 6 sites and across time; for the latter, they conducted an ANOVA to determine whether changes over time were statistically significant. This can be considered a weak form of construct validation (i.e., comparing the score with itself across time).
    - Although the ***overall method*** (ANOVA) is appropriate, the developers appear to have aggregated data at two levels to conduct the ANOVAs (e.g., for providers, then by site), and therefore the results do not seem to demonstrate differences for the four levels of analysis. Moreover, additional testing would be needed before conferring endorsement for all four levels of analysis, as only two sets of results were provided. NQF staff have requested additional information from the developer, but it is not yet available.

- Face validity
    - The developer provides thoughtful and detailed discussion about several components of the measure. While informative and useful to consider for validity overall, this qualitative analysis does not meet NQF's testing requirements for assessing the face validity of the measure score as an indicator of quality.
- Additional analyses
    - For 225 encounters (a stratified random sample of 100 from eligible visits, 100 from non-eligible primary care visits, and 25 from non-eligible otorhinolaryngology visits) from one site, the developer manually compared the eMeasure results to results obtained when an abstractionist used both discrete EHR fields as well as clinical notes.
        - While this analysis does not meet NQF's criteria for data element testing (as individual data elements were not compared and data from only one site were examined), it does speak somewhat to quality of the data in the defined fields.
    - For one test site, developers correlated measure results found if using ICD-9-CM coding versus that used if using a proprietary coding system.

Validity testing results:

- Score-level testing: Because the ANOVAs appear to be conducted on aggregated data rather than for each level of analysis as specified, additional information from the developer is required before the results can be interpreted.
- Additional analysis: comparison to result from manual abstraction: sensitivity=0.90; specificity=0.92
- Additional analysis: ICD-9-coding vs. proprietary coding: The measure results were strongly correlated, regardless of which coding system was used.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- ***If*** *the aggregation for the ANOVA is appropriate, do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you have any concerns about the validity of the measure as specified?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The denominator exclusions include diagnoses at the visit of common childhood infection for which antibiotics are frequently indicated. Developers state that these diagnoses were *"… chosen based on established prevalence of childhood infections, as well as analysis of most common diagnoses co-occurring with antibiotic prescription in a large pediatric care network*."
- The exclusions analysis is somewhat unclear, but it appears that the developers provided a sensitivity analysis that compares the measure scores at the site level with and without applying the exclusions specified in the measure. The results indicate substantial differences in measure score if exclusions are not applied.

Site	% abx, excluded enc	% abx, all enc
A	10	13
B	35	13

C	32	23
D	53	33
E	4	2
F	11	3

Questions for the Committee:

- *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)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Developers provided pictorial representation of the variations between clinicians and practices and over time.
- Although the ANOVA results are not precisely what is needed for reliability and validity testing, the distributional data demonstrate variability in results between clinicians and practices.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable.

2b7. Missing Data

- Developers presented the extent of missing data, based on information from PEDSnet (this includes the 6 sites used in the other analyses presented in the submission.

Across PEDSnet, data element feasibility for the Fall 2016 data cycle are:

Data Element	Missingness (%)	Comment
Patient date of birth	0	
Encounter date	0	
Encounter type	0	Outpatient/ED/Inpatient
Diagnosis standard code	0.1	
Diagnosis-encounter link	1.5	Omitting problem list entries
Medication standard code	10.5	Source (vendor) code 0% missing
Medication-visit link	0.6	
Encounter provider ID	0.1	Used for testing; not required for measure computation

Question for the Committee:

- *Do you understand the 10.5% missing value for the Medication standard code element? Is it of concern?*

Guidance from the Validity Algorithm
Specifications are consistent with the evidence (Box 1) → Potential threats to validity were assessed (Box 2) → Empirical testing was conducted, but the testing does not appear to match the measure specifications (Box 3) → No face validity assessment for the measure as an indicator of quality, per NQF requirements (Box 4) → Insufficient

The highest possible rating is HIGH.

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

RATIONALE: It appears that the method used for testing is not quite appropriate, but additional information from the developer is needed. Also, additional testing will be needed in order to endorse the measure for all four levels of analysis specified.

Committee pre-evaluation comments

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

2a1 & 2b1

- Is the test sample adequate to generalize for widespread implementation? Yes
- If the aggregation for the ANOVA is appropriate, do the results demonstrate sufficient reliability so that conclusions about quality can be made? Because the ANOVAs appear to be conducted on aggregated data rather than for each level of analysis as specified, additional information from the developer is required before the results can be interpreted.
- Six academic pediatric centers were included in testing, including the measure development site. All institutions provide secondary through quaternary care, including multiple specialty outpatient practices and pediatric emergency departments. Outpatient primary care facilities ranged from large regional networks to hospital clinics providing primary care to patients otherwise engaged in specialty care at the institution, with the availability of specific services varying by institution.
- For each institution, all available outpatient or emergency department settings, and all providers with encounters in these settings, were included in testing spanning the entire time interval.
- All but one site got data from community docs.
- Data elements outlined are not just OME. Could not discern which elements would be used for inclusion versus exclusion when an antibiotic is prescribed.
- Measure depends on accurate diagnosis. Pichichero, 2001: The distinction between acute suppurative otitis media (AOM) and otitis media with effusion (OME) is important for antibiotic treatment decisions. Overall, the average correct diagnosis by pediatricians was 50% (range, 25%-73%) and by otolaryngologists was 73% (range, 48%-88%). http://jamanetwork.com/journals/jamapediatrics/fullarticle/191139
- Data elements are clearly defined. Appropriate codes are included. The measure can likely be consistently implemented. The specifications are consistent with the evidence.

2b.2

- Most significant issue is that doc may be prescribing appropriately for another indication.
- Also it is noted, "It appears that the method is not quite appropriate, but additional information from the developer is needed. Also, additional testing will be needed in order to endorse the measure for all four levels of analysis specified. If the method is not appropriate, the rating for the criteria would be LOW, unless the developer can provide additional analysis. "
- Insufficient testing sites. This is not adequate to generalize for widespread implementation.
- The reliability was again insufficient.
- Missing complete analysis. Unable to determine if measure is precise.
- The testing sample is adequate but could include more data on more recent patient encounters as practice patterns could be evolving over time.

2b.2

- Presenting data are from only one site and is not representative of the US and territories.
- The validity appeared adequate.
- Missing complete analysis.
- The patient population examined is generally valid but as some of the data is several years old and practice patterns are evolving it might not be as relevant as needed. Collection os data from the measure as it is implemented may help to clarify.

2b3-7

- I do not understand the threat to validity table. Also do I do not understand what they are talking about on page 14 of OME testing pdf attachment "Mean scores were high, but appreciably different from ideal, and lower than would be expected solely from the prevalence of alternate indications for antibiotics not measured, indicating that current practice has opportunity in many cases for measurable improvement. Conversely, entities do reach scores >0.95, and the third quartile for several sites reaches 1.00 in data through 2016, confirming that it is possible to reach very high measure scores in practice; the effective “ceiling” set by gaps in data capture does not greatly limit the dynamic range of the measure. This is true for clinical specialties ranging from primary care, where most children with OME are seen, to otorhinolaryngology, where more complex cases are evaluated; the higher mean scores for the latter case may imply that greater experience correlates with higher measure scores,

similar to results seen for all entities with higher evaluable visit counts." Also the table on page 15 is not clear to me. But does not appear to be much missing data and exclusions make a lot of sense.
- Exclusions were not clearly defined. Risk adjustment was not addressed adequately. This measure did not address meaningful differences about quality.
- There was a large range in the sites based on exclusions. This needs clarification. Sites E and F had no gap in care. Were some sites tertiary centers treating sicker children?
- For the missing medication standard code, would this not be missing frequently since we are looking at medication avoidance?
- The exclusions of other conditions that might require antibiotics are consistent with the evidence. There are no patient groups inappropriately excluded. Yes the measure identifies meaningful differences about quality.

Criterion 3. Feasibility

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.

- The data elements are available in electronic health records (EHR) or other electronic sources.
- This is an eMeasure and a feasibility scorecard was provided by the measure developer.
- All required data elements appear to be in defined fields in EHRs.
- The feasibility scorecard addresses the main components of feasibility, but it is not clear to which EHR/site the scorecard reflects (it is possible that it is reflective of all 6 sites tested, but this should be clarified).

Questions for the Committee:

- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

- All of the data elements should be routinely generated during care. The required data elements should be available in electronic form.
- The feasibility appears adequate.
- It is assumed that all entities use EHR or electronic sources and that the data elements are already built to be easily retrieved from an electronic data warehouse. This assumption has not been proven and raises concerns about the data collection strategy feasibility.
- It is relatively easy to get the data the way they have arranged and asked for it. The primary concern is that it relies on coding and as they mention maybe OME will be 2nd or 3rd code and first will be indication for abx.

Criterion 4: Usability and Use

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

4. Usability and 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- The developer indicates that the intended use of the measure is to be included in a public reporting program for children enrolled in Medicaid and CHIP.

Improvement results: N/A (this is a new measure)

Unexpected findings (positive or negative) during implementation : None reported.

Potential harms: None identified.

Vetting of the measure: None reported.

Feedback: N/A

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

- The related measures work well with this one and all have the same format of data collection. Could truly consider figured out a way that all 3 other measures were combined. Clearly the measure can be used to further high-quality care as it will hopefully result in less abx usage. The main worry with all abx checking methods is that docs still write abx just code better, so would be nice to consider how to collect that data.
- Could not identify how this measure is being publicly reported. Want to be sure that the measure does not penalize a professional for prescribing antibiotics when necessary due to other identified bacterial infections not listed as exclusions. Would like to explore the feasibility of claims data in place of data extraction from an EHR.
- The measure is new and currently not being used for public reporting.
- The results can be used to further the goal of high quality care by providing feedback to clinicians on the quality of their care and by tying reimbursement to the quality delivered. I do not see potential unintended consequences.

Criterion 5: Related and Competing Measures

Related measures

- 0655 : Otitis Media with Effusion: Antihistamines or decongestants – Avoidance of inappropriate use
- 0656 : Otitis Media with Effusion: Systemic corticosteroids – Avoidance of inappropriate use
- 0657 : Otitis Media with Effusion: Systemic antimicrobials – Avoidance of inappropriate use

Harmonization

- The Committee may be asked to discuss whether there are any facets of the measures that should be harmonized.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as

demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes X No**

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because it has not been well-vetted in real world settings by those being measured and other users.

Pre-meeting public and member comments

Comment by American Academy of Otolaryngology – Head and Neck Surgery: While the exceptions are somewhat different, #2640 Otitis Media with Effusion - Antibiotics Avoidance e-measure is very similar to NQF 0657 - Otitis Media with Effusion: Systemic Antimicrobials - Avoidance of Inappropriate Use, which has already been endorsed, and is stewarded by the AAO-HNSF. The AAO-HNSF encourages harmonization between these two measures before #2640 is further considered for endorsement.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

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

Measure Title: Otitis Media with Effusion (OME) Antibiotic Avoidance Measure

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 1/18/2017

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Avoidance of inappropriate antibiotics usage

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

NA

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

For this measure, the health outcome we measure is the appropriate use of antibiotics, specifically the avoidance of antibiotics to treat Otitis Media with Effusion. Systemic antimicrobial therapy does not result in clinical improvement of OME, as it is not an infectious disease. The measure aims to minimize the use of ineffective treatment with multiple adverse consequences, including increased prevalence of antibiotic-resistant flora, increased cost, and increased side effect burden (*e.g.* antibiotic-associated diarrhea, drug allergy). Since antibiotics are available only by prescription from an appropriately licensed clinician in the United States, a process measure reporting on prescribing practice associated with OME diagnosis is a valid proxy for antibiotic usage associated with OME.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

American Academy of Family, P., American Academy of, Otolaryngology-Head and Neck, Surgery and American Academy of Pediatrics Subcommittee on Otitis Media With Effusion (2004). Otitis media with effusion. Pediatrics *113*, 1412-1429.

Rosenfeld RM, Shin JJ, Schwartz SR, Coggins R, Gagnon L, Hackell JM, Hoelting D, Hunter LL, Kummer AW, Payne SC, Poe DS, Veling M, Vila PM, Walsh SA, Corrigan MD. Clinical Practice Guideline: Otitis Media with Effusion (Update). Otolaryngol Head Neck Surg. 2016 Feb;154(1 Suppl):S1-S41.

http://oto.sagepub.com/content/154/1_suppl/S1.full

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Rosenfeld RM, Shin JJ, Schwartz SR, Coggins R, Gagnon L, Hackell JM, Hoelting D, Hunter LL, Kummer AW, Payne SC, Poe DS, Veling M, Vila PM, Walsh SA, Corrigan MD. Clinical Practice Guideline: Otitis Media with Effusion (Update). Otolaryngol Head Neck Surg. 2016 Feb;154(1 Suppl):S1-S41.

Page S20

"STATEMENT 8b. ANTIBIOTICS: Clinicians should recommend against using systemic antibiotics for treating OME. Strong recommendation against based on systematic review of RCTs and preponderance of harm over benefit."

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

Strong recommendation	A strong recommendation means that the benefits of the recommended approach clearly exceed the harms (or, in the case of a strong negative recommendation, that the harms clearly exceed the benefits) and that the quality of the supporting evidence is high (grade A or B). In some clearly identified circumstances, strong recommendations may be made based on lesser evidence when high-quality evidence is impossible to obtain and the anticipated benefits strongly outweigh the harms.	Clinicians should follow a strong recommendation unless a clear and compelling rationale for an alternative approach is present.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Table 4. Strength of Action Terms in Guideline Statements and Implied Levels of Obligation.

Strength	Definition[a]	Implied Obligation
Strong recommendation	A strong recommendation means that the benefits of the recommended approach clearly exceed the harms (or, in the case of a strong negative recommendation, that the harms clearly exceed the benefits) and that the quality of the supporting evidence is high (grade A or B). In some clearly identified circumstances, strong recommendations may be made based on lesser evidence when high-quality evidence is impossible to obtain and the anticipated benefits strongly outweigh the harms.	Clinicians should follow a strong recommendation unless a clear and compelling rationale for an alternative approach is present.
Recommendation	A recommendation means that the benefits exceed the harms (or, in the case of a negative recommendation, that the harms exceed the benefits), but the quality of evidence is not as high (grade B or C). In some clearly identified circumstances, recommendations may be made based on lesser evidence when high-quality evidence is impossible to obtain and the anticipated benefits outweigh the harms.	Clinicians should also generally follow a recommendation but remain alert to new information and sensitive to patient preferences and modifying factors.
Option	An option means that either the quality of evidence is suspect (grade D) or that well-done studies (grade A, B, or C) show little clear advantage to one approach versus another.	Clinicians should be flexible in their decision making regarding appropriate practice, although they may set bounds on alternatives; patient preference should have a substantial influencing role.

[a]See Table 5 for definitions of evidence grades.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Rosenfeld RM, Shiffman RN, Robertson P, et al. *Clinical practice guideline development manual, third edition: a quality-driven approach for translating evidence into action. Otolaryngol Head Neck Surg. 2013;148(1):S1-S55.*

http://oto.sagepub.com/cgi/ijlink?linkType=ABST&journalCode=spoto&resid=148/1_suppl/S1

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

NA

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

NA

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

NA

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

NA

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

NA

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

NA

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The avoidance of systemic antibiotics for treating OME

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Table 5. Aggregate Grades of Evidence by Question Type.[62]

Grade	Treatment	Diagnosis	Prognosis
A	Systematic review[a] of randomized trials	Systematic review[a] of cross-sectional studies with consistently applied reference standard and blinding	Systematic review[a] of inception cohort studies[b]

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Table 5. Aggregate Grades of Evidence by Question Type.[62]

Grade	Treatment	Diagnosis	Prognosis
A	Systematic review[a] of randomized trials	Systematic review[a] of cross-sectional studies with consistently applied reference standard and blinding	Systematic review[a] of inception cohort studies[b]
B	Randomized trials or observational studies with dramatic effects or highly consistent evidence	Cross-sectional studies with consistently applied reference standard and blinding	Inception cohort studies[b]
C	Nonrandomized or historically controlled studies, including case-control and observational studies	Nonconsecutive studies, case-control studies, or studies with poor, nonindependent, or inconsistently applied reference standards	Cohort study, control arm of a randomized trial, case series, or case-control studies; poor quality prognostic cohort study
D	Case reports, mechanism-based reasoning, or reasoning from first principles		
X	Exceptional situations where validating studies cannot be performed and there is a clear preponderance of benefit over harm		

[a]A systematic review may be downgraded to level B because of study limitations, heterogeneity, or imprecision.
[b]A group of individuals identified for subsequent study at an early uniform point in the course of the specified health condition or before the condition develops.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: through Jan 2015

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

20 Systematic reviews

49 Randomized controlled trials

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The aggregate evidence quality was assessed as Grade A, given a systematic review of well-designed RCTs. The level of confidence in the data was high, with little vagueness, resulting in no differences in opinion among expert reviews. The result was a strong recommendation for the measure use.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

The estimated benefits include the avoidance of side effects and reduction in cost by not administering medications; avoidance in delays in definitive therapy caused by short-term improvement then relapse; and avoidance of societal impact of inappropriate antibiotic prescribing on bacterial resistance and transmission of resistant pathogens. The preponderance of evidence suggests benefit over harm in the recommendation against therapy.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

"side effects and reduction in cost by not administering medications; avoidance in delays in definitive therapy caused by short-term improvement then relapse; and avoidance of societal impact of inappropriate antibiotic prescribing on bacterial resistance and transmission of resistant pathogens"

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

NA

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

NA

1a.8.2. Provide the citation and summary for each piece of evidence.

NA

1. Evidence, Performance Gap, Priority – Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***
1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form 3._OME_MeasSubm_Evidence_FINAL.docx
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., the benefits or improvements in quality envisioned by use of this measure)*
Otitis media is a highly prevalent condition among young children particularly, and a major driver of outpatient health care and antibiotic utilization. Otitis Media with Effusion (OME), a non-infectious chronic condition, occurs in 6.5% of children 0-11 years of age in our test population; it is well-established that, unlike acute otitis media, antibiotic therapy provides no benefit for OME. Nonetheless, antibiotics have been frequently prescribed in this context. Reduction of inappropriate antibiotic use has significant public health benefits, including reduced development of antibiotic resistance, reduced health care cost, and decreased side effect burden for patients.

This measure is developed for evaluation in electronic health record data, providing quality measurement across a large range of population sizes and data types. The existing Academy of Otolaryngology-Head and Neck Surgery (AAOHNS)/Academy of Family Physician´s (AAFP) PCPI measures are specified as manual chart review measures, limiting sample size for evaluation. They also used uncommon coding systems, which was the proximate cause of the failure of the initial CHIPRA core set OME measure. We are addressing both issues, and testing with data from multiple EHR systems.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The measure was tested on 36,060 encounter records drawn from CHOP EHR spanning 2009-2014. For provider-specific evaluation (N=531; with 285 representing providers with 5 or more visits), score characteristics were as follows:
mean failure rate 15.05%; median 8.00%; IQR 3.00% - 20.00%. For specialty-specific evaluation (N=3), score characteristics were as follows: mean failure rate 11.42%; median 6.00%; IQR 0.00% - 16.00%. We examined General Pediatrics, Otorhinolaryngology, and Other departments as a group.

The analysis was expanded to include data from six health systems using a total of three EHR systems (Epic Systems EHR, Cerner's Millennium EHR, or Allscripts' EHR). More detailed data regarding performance are available in the measure testing form.

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.
Performance data provided in 1b2.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Race/Ethnicity

Race and ethnicity was determined based on demographic data record in the EHR. To test for racial/ethnic disparities, the measure was evaluated at the provider level, stratified by race/ethnicity (Black non-Hispanic, Hispanic, White non-Hispanic, Other). Chi-squared testing identified statistically significant differences in measure scores across groups because of the large sample size, but the size of the differences detected were small.
Socioeconomic Status
Socioeconomic status at the individual level was examined using insurance information from the EHR, categorized into three groups: ever reported public insurer, reported only commercial insurer, and no reported insurer. Again, Chi-squared testing identified statistically significant differences in measure scores across groups because of the large sample size, but the size of the differences detected were small.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
Performance data provided in 1b.4

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

- Approximately 2.2 million new cases of OME are diagnosed annually in the United States, with 50% to 90% of diagnoses made by 5 years of age.
- The point prevalence is 7% to 13%, with a peak in the first year of life and a per-year period prevalence of 15% to 30%.
- The annual cost of care for OME in the United States has been estimated at $4.0 billion.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Rosenfeld RM, Culpepper L, Doyle KJ, et al. Clinical practice guideline: otitis media with effusion. Otolaryngol Head Neck Surg. 2004;130(5):S95-S118.
2. Tos M. Epidemiology and natural history of secretory otitis. Am J Otol. 1984;5:459-462.
3. Casselbrant ML, Mandel EM. Epidemiology. In: Rosenfeld RM, Bluestone CD , eds. Evidence-Based Otitis Media. 2nd ed. Hamilton, Canada: BC Decker Inc; 2003:147-162.
4. Zielhuis GA, Rach GH, van den Broek P. Screening for otitis media with effusion in preschool children. Lancet. 1989;1:311-314.
5. Casselbrant ML, Brostoff LM, Cantekin EI, et al. Otitis media with effusion in preschool children. Laryngoscope. 1985;95:428-436.
6. Martines F, Martines E, Sciacca V, Bentivegna D. Otitis media with effusion with or without atopy: audiological findings on primary schoolchildren. Am J Otolaryngol. 2011;32:601-606.
7. Mandel EM, Doyle WJ, Winther B, Alper CM. The incidence, prevalence and burden of OM in unselected children aged 1-8 years followed by weekly otoscopy through the “common cold” season. Int J Pediatr Otorhinolaryngol. 2008;72:491-499

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not applicable

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 subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

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

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

Ears, Nose, Throat (ENT) : Ear Infection

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

«crosscutting_area»

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.)*
Not applicable

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 an eMeasure **Attachment:** OMEAvoidance_v4_6_Artifacts_-1-.zip

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:** OME_VSAC_ValueSets.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Not applicable

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Eligible encounters at which a systemic antibiotic was not prescribed.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The measure may be evaluated over any interval that allows the accumulation of a sufficient number of eligible visits (i.e. >5 eligible encounters per evaluated entity). The measurement time period may vary upon needs of the particular user (e.g. calendar year, quarterly, monthly), but must be the same for both the numerator and denominator.

S.6. **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, 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.
Encounters meeting eligibility criteria (see denominator statement) at which there is no record of a systemic antibacterial antibiotic prescription.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Outpatient encounters at which otitis media with effusion is diagnosed, but at which common conditions for which antibiotics are indicated are not diagnosed. It is expected that a small fraction of patients with rare non-OME indications for antibiotic usage will not be identified by the specified exclusion criteria, but these will be rare cases, and will not alter the measure score significantly in most practice contexts. Of note, however, applicability may be limited in specific practice environments in which a large proportion of patients seen have immune deficiencies requiring chronic antibiotic use (e.g. immunology or hematology/oncology clinics).

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Outpatient encounters (including office/clinic, emergency department, and urgent care but not including ambulatory surgery centers) meeting all of the following criteria:

1. Patient age two months through 155 months, inclusive, on date of visit;

2. Outpatient encounter(s), using criteria appropriate to the source system. These criteria may include system-specific encounter type codes, department or clinic identifiers, or E&M codes indicative of outpatient clinical service (e.g. where CPT4 codes are used to define encounter types, the following list might be included: 99201, 99202, 99203, 99204, 99205, 99211, 99212, 99213, 99214, 99215, 99241, 99242, 99243, 99244, 99255, 99281, 99282, 99283, 99284, 99285, 99381, 99382, 99383, 99384, 99391, 99392, 99393, 99394);
3. Diagnosis at the visit of Otitis Media with Effusion (OME), as specified in the value set noted above for systems using ICD-9-CM, ICD-10-CM, or SNOMED-CT as their diagnostic terminology;
4. Medication prescribing data available in the source system (this is a system requirement; no medications need have been prescribed at the specific visit)

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Diagnosis at the visit of common childhood infection for which antibiotics are frequently indicated.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
Diagnosis of common infection for which antibiotics are frequently indicated, as specified in the value set noted above for source systems using ICD-9-CM, ICD-10-CM, or SNOMED-CT as their diagnostic terminology. These codes were chosen based on established prevalence of childhood infections, as well as analysis of most common diagnoses co-occurring with antibiotic prescription in a large pediatric care network.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Measure validity is not dependent on stratification, but an organization may consider stratifying by sociodemographic factors in order to assess disparities in care provide in Otitis Media with Effusion.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not Applicable

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
1. Specify the desired time period for evaluation
2. Identify all eligible denominator encounters within the specified time period for the entity being measured;

3. For each encounter in the denominator set, add to the numerator if the encounter meets numerator inclusion criteria;
4. Compute (count of numerator encounters) / (count of denominator encounters)

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable, all available records are used.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Not Applicable

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Other

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Electronic Health Records

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Facility, Integrated Delivery System

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Clinician Office/Clinic, Urgent Care - Ambulatory
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
4._OME_testing_attachment_2017_FINAL-636203662987273712.pdf

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Otitis Media with Effusion Antibiotic Avoidance Measure
Date of Submission: 1/18/2017
Type of Measure:

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

Instructions

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

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

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

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

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

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Computable eMeasure; EHR does not implement HQMF	☒ other: Computable eMeasure; EHR does not implement HQMF

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*).

Computable measure testing was performed on clinical data extracted from the EHR systems of six pediatric health systems, using three different EHR vendors (Epic Systems EHR, Cerner's Millennium EHR, and a combination of Allscripts' EHR and an institutional Emergency Department system). The dataset included all available coded diagnoses, encounters, and medication utilization data for all patients seen in each health system since 2009. Prior to testing, data from each EHR was standardized to the PEDSnet Common Data Model v2.3, derived from the Observational Medical Outcomes Partnership Common Data Model v5. Standard terminologies used within the CDM were drawn from the OMOP Vocabulary service in November 2016.

1.3. What are the dates of the data used in testing?

Aggregate testing was performed on records from 2009-01-01 through 2016-06-30. Since deployment and upgrades to the EHR were regularly performed at testing sites through this interval, year-over-year comparisons were restricted to the interval 2012-01-01 through 2016-06-30, during which the scale of these changes was significantly reduced relative to earlier periods.

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan

☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Six academic pediatric centers were included in testing, including the measure development site. All institutions provide secondary through quaternary care, including multiple specialty outpatient practices and pediatric emergency departments. Outpatient primary care facilities ranged from large regional networks to hospital clinics providing primary care to patients otherwise engaged in specialty care at the institution, with the availability of specific services varying by institution.

For each institution, all available[1] outpatient (non-ASF) or emergency department settings, and all providers with encounters in these settings, were included in testing spanning the entire time interval.

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

All patients with at least one evaluable visit were included in testing.

Site	N	Male %	White %	Black %	Other race %	0-2 y %	2-5 y %	6-12 y %
A	19070	57.49	61.63	23.29	15.08	41.86	40.77	17.34
B	30713	58.33	62.41	17.14	20.45	31.07	47.73	21.20
C	11726	58.18	65.82	5.09	29.09	36.98	43.46	19.56
D	29258	58.35	69.42	12.71	17.87	37.12	42.04	20.83
E	6079	59.16	77.40	15.92	6.68	38.95	43.66	17.39
F	7803	58.10	59.87	6.33	33.80	34.74	41.59	23.66

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.

For year-over-year analyses, inclusion was limited to entities having ≥5 evaluable events in each year, in order to limit the impact of hysteresis due to limited evaluable events per entity.

Individual chart abstraction was done using the EHR user interface at the Children's Hospital of Philadelphia. The sampling strategy for manual record review was as follows:

- 100 encounters were randomly selected from eligible visits (*i.e.* from the baseline population after all exclusion criteria are applied), insuring proportional representation of departments accounting for ≥5% of visits. At the provider level, a 2-fold oversampling was done of the 10% of providers with the highest and lowest failure rates, considering only providers that have ≥5 eligible visits, and requiring 75% of overall visits to come from providers with ≥5 eligible visits.

[1] Site E reports incomplete capture of data from affiliated outpatient practices not fully owned by the institution

- One hundred (100) encounters were randomly selected from non-eligible primary care encounters, and 25 were selected from non-eligible otorhinolaryngology (ENT) encounters, as these two clinical departments accounted for the majority of eligible encounters.

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

None

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*)

This is a computable eMeasure, specified using

- extensional value sets
- data elements that result from specified positive steps in the clinical care process, and are not imputed or estimated, and
- an algebraic scoring rule.

As such, it is susceptible to, and was tested using, a deterministic imperative algorithm. Because this approach guarantees identical results on repeated evaluation of a given set of input data, traditional reliability testing of the measure scoring process was not performed.

Comparison of computed values of critical data elements to chart abstraction, and of measure evaluation across multiple entities, was performed as part of validity testing, and provides indirect information about stability across evaluation context. While this does not meet the formal definition of reliability, given different methods (data elements) and clinical differences across entities (measure score), this information may nonetheless be useful to the evaluator in assessing the variance of the measure across situations likely to be encountered in practice.

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*)

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?*)

2b2. VALIDITY TESTING

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

☒ **Performance measure score**

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

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

Data elements**:** Feasibility of critical data elements for measure computation is assessed as part of PEDSnet routine data quality analysis, spanning eight member sites and three EHR vendors, including the six sites participating in testing of this measure.

Measure score, face validity: Components of the measure specification were compared to the AAP/AAFP 2004 and AAO-HNS 2016 OME treatment guidelines to assess the congruence of measure components to the guidelines, and the correlation between higher measure score and care consistent with the recommendation that antibiotics are not an effective therapy for OME. Results of the assessment are reported qualitatively.

Measure score, construct validity: Accuracy of measure components derived from EHR data elements was assessed by manual review of 225 patient records, using the sampling strategy described above. Review included both discrete data and clinician notes. Results were compared to the measure classification assigned by the computable specification, and reported as sensitivity and specificity relative to the benchmark manual review.

Measure score, concurrent validity: The measure was computed over the population of patients seen 2009-2016 at six pediatric institutions, and success rates were summarized by provider and by clinical department. For one institution, the dataset received contained gaps in RxNorm code assignment; consultation with the institution revealed that this was due to a gap in mapping between the vendor codes used operationally and equivalent RxNorm codes. As this was a technical problem expected to be temporary, measure testing was done using generic drug names for this site, in order to permit assessment of other measure components. One site reported inability to collect data from affiliated community practices, so testing was limited to care delived at the children's hospital; this did not compromise the technical validity of available data, but resulted in a smaller proportion of eligible encounters than at other sites. Scores were compared across institutions to examine performance in different contexts, as well as to assess the range and variability of results. Scores for designated time periods within 2012-16 were also computed at each institution for a fixed set of entities, in order to estimate change over time per provider or department. Data are reported as success rate per entity, with descriptive statistics. ANOVA is used to test significance of differences.

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

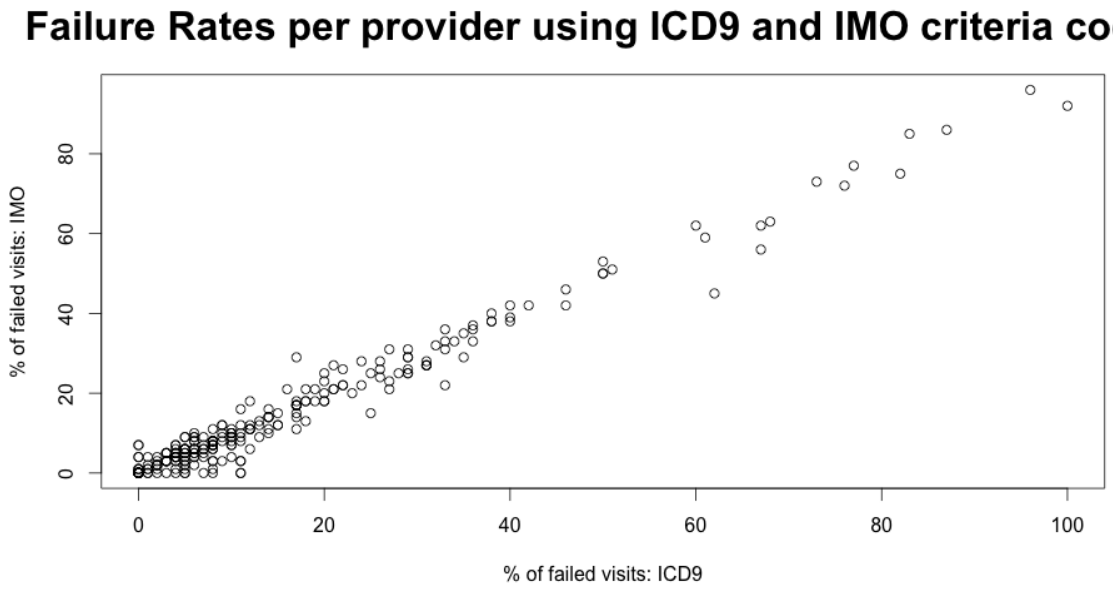
Data elements: Across PEDSnet, data element feasibility for the Fall 2016 data cycle are:

Data Element	Missingness (%)	Comment
Patient date of birth	0	
Encounter date	0	
Encounter type	0	Outpatient/ED/Inpatient
Diagnosis standard code	0.1	
Diagnosis-encounter link	1.5	Omitting problem list entries
Medication standard code	10.5	Source (vendor) code 0% missing

Medication-visit link	0.6	
Encounter provider ID	0.1	Used for testing; not required for measure computation

Measure score, face validity: The measure specification is aligned with evidence-based recommendation for the avoidance of antibiotics in several respects:

- *Coded diagnosis of OME* – The diagnosis of OME relies on the presence and chronicity of middle ear effusions, and can be complex and difficult to ascertain independently from the medical record. Since the intent of the measure is to assess appropriateness of treatment for OME, the specification requires a coded diagnosis of OME as evidence of the clinician's diagnostic assessment and involvement in medical decision-making. This improves specificity by decreasing potential for over-ascertainment, particularly of borderline cases, by attempts to parse physical exam findings or similar narrative data. It does create the risk that OME as a secondary problem at a visit will not be recorded. This may bias evaluation toward a lower measure score, since a diagnosis is less likely to be omitted if it drives a decision to prescribe antibiotics, whether appropriate or not. There is also the potential that a clinician uses an OME code for a diagnosis intermediate between OME and AOM, or uses a nonspecific otitis media code; to the extent that the antibiotic prescribing decision is driven by acuity of the otitis, the measure specification may drive increased accuracy of coding. Finally, the difference between the (often proprietary or site-specific) terminology presented to the clinician in the EHR and the (standard) coding system to which it is mapped, such as ICD9-CM or SNOMED-CT, creates the potential for misclassification due to erroneous mapping. We assessed this risk by scoring the measure using both the ICD9-CM standard and the proprietary IMO terminologies at one test site, and noted strong correlation:



- *Age range* – As the clinical guideline notes, most cases of OME occur prior to 5 years of age. As age increases, the likelihood that a diagnosis of OME represents different pathophysiology than in younger children increases. It is not clear that antibiotic treatment is more appropriate in such cases, but the small number of overall diagnoses, and the increased likelihood of a coding error, suggest that extending the age range beyond that currently used for the analogous chart review measure may increase misclassification with minimal increase in quality assessment.
- *Antibiotic prescription* – While no type of antibiotic treatment is indicated for OME, the use of systemic antibiotics is expected to have greater burden of adverse effects (*e.g.* diarrhea, systemic allergy) and a greater impact on antibiotic resistance than topical antibiotics. The measure is therefore specified to address systemic antibiotic prescription specifically. Because medication data is required for evaluation, it is expected that the measure will only be useful in the EHR context when computerized provider order

entry has been implemented; the specification does not contemplate indirect ascertainment of antibiotic prescription from narrative such as clinician notes or patient instructions. However, CPOE is recognized as a significant benefit of EHR use, and the CMS EHR Incentive program Stage 1 focused strongly on adoption of CPOE, so it is expected that available prescription data will reflect actual antibiotic usage, and therefore that the measure will address antibiotic avoidance as an indicator of appropriate care for OME. Ability to evaluate antibiotic usage also depends on the coding system used by the EHR, since the drug value set is specified using RxNorm. This terminology was chosen both for its semantics, in part because of its standard usage in the OMOP CDM used for testing, and because the CMS EHR Incentive Program Stage 2 includes interoperability requirement that incorporate RxNorm as the standard drug terminology, increasing the likelihood that EHR vendors and health systems will support transformation to this terminology.

- *Concurrent diagnoses* – Perhaps the greatest threat to validity of the measure is that possibility that antibiotics prescribed concurrently with a diagnosis of OME do not reflect an inappropriate treatment decision for OME, but the presence of an alternate indication for antibiotics. The number of potential diagnosis codes for either constitutional or acute conditions requiring antibiotics is quite large, but the majority of these conditions are rare in the population. In these cases, comprehensively enumerating such codes would greatly increase the complexity of measure maintenance, while having a marginal impact on the ability of the measure to report on quality of decision-making in the typical case. The measure specification attempts to balance these concerns by incorporating those infections we found co-occurred most commonly with antibiotics in our test cohort: acute otitis media, sinusitis, pneumonia, and pharyngitis. In the latter two cases, we have included the diagnostic codes for the illness without a specified pathogen, in recognition that clinicians may use a less specific code in advance of or in the absence of laboratory results. While this may lead to unnecessary antibiotic use for other conditions, that is properly the focus for other measures, and does not reflect the quality of decision-making regarding OME.
- *Measure score* – The score computation for the measure reflects in a straightforward way the desired quality outcome: higher scores indicate a larger proportion of OME cases in which antibiotics were avoided.

Measure score, construct validity: Sensitivity: 0.90 Specificity: 0.92

Measure score, concurrent validity: Measure computation by provider across all care contexts at test sites for the 2009-2016 interval, among providers with ≥5 evaluable visits in a given year, produced the following score distributions:

Site	Min	1st Q	Mean	SD	3rd Q	Max
A	0.00	66.67	79.83	24.81	100.00	100.00
B	0.00	55.56	74.16	30.29	100.00	100.00
C	0.00	53.95	71.72	33.50	97.82	100.00
D	14.29	100.00	95.94	12.98	100.00	100.00
E	0.00	98.94	96.09	13.67	100.00	100.00
F	0.00	50.00	65.01	25.01	83.33	100.00

with ANOVA yielding an F statistic 55.68 ($p<0.001$). When aggregated by clinical department, score distributions were:

Site	Min	1st Q	Mean	SD	3rd Q	Max
A	33.33	75.00	81.91	18.20	96.61	100.00
B	33.33	66.30	81.32	20.00	100.00	100.00
C	39.02	66.67	80.17	18.86	97.44	100.00
D	28.57	99.47	96.45	10.23	100.00	100.00
E	80.00	93.33	95.84	06.08	100.00	100.00

F	37.56	65.17	76.00	15.19	85.71	100.00

with ANOVA yielding an F statistic 16.47 ($p<0.001$).

Annual measure computation for the subset of providers having ≥5 evaluable encounters in each year yielded:

Site	2012	2013	2014	2015	2016[2]	F	p
A	84.66	84.09	76.90	58.67	65.93	28.01	<.001
B	80.73	82.64	80.65	77.81	74.19	1.359	.245
C	81.14	79.69	83.70	85.43	87.52	1.016	.317
D	75.18	69.85	67.24	56.75	58.92	21.88	<.001
E	99.57	98.20	99.37	99.31	99.90	2.592	.114
F	97.05	96.14	96.07	97.45	95.16	.100	.752

A similar analysis aggregated by clinical department yielded:

Site	2012	2013	2014	2015	2016[3]	F	p
A	90.84	89.57	88.76	72.50	71.12	12.27	.001
B	85.83	87.86	79.10	73.20	76.24	3.462	.068
C	70.58	72.92	75.28	77.74	69.10	0.006	.939
D	77.70	71.00	73.19	63.54	70.33	2.563	.117
E	96.75	95.96	97.37	92.16	96.71	.188	.670
F	94.55	94.45	96.40	96.37	91.35	.167	.686

Finally, because the diagnostic coding system for billing processes in the U.S. changed in October 2015 from ICD-9-CM to ICD-10-CM, and this may have led to changes in point-of-care diagnosis coding, we examined results for the 9 months prior and the 9 months following 2015-Oct-01:

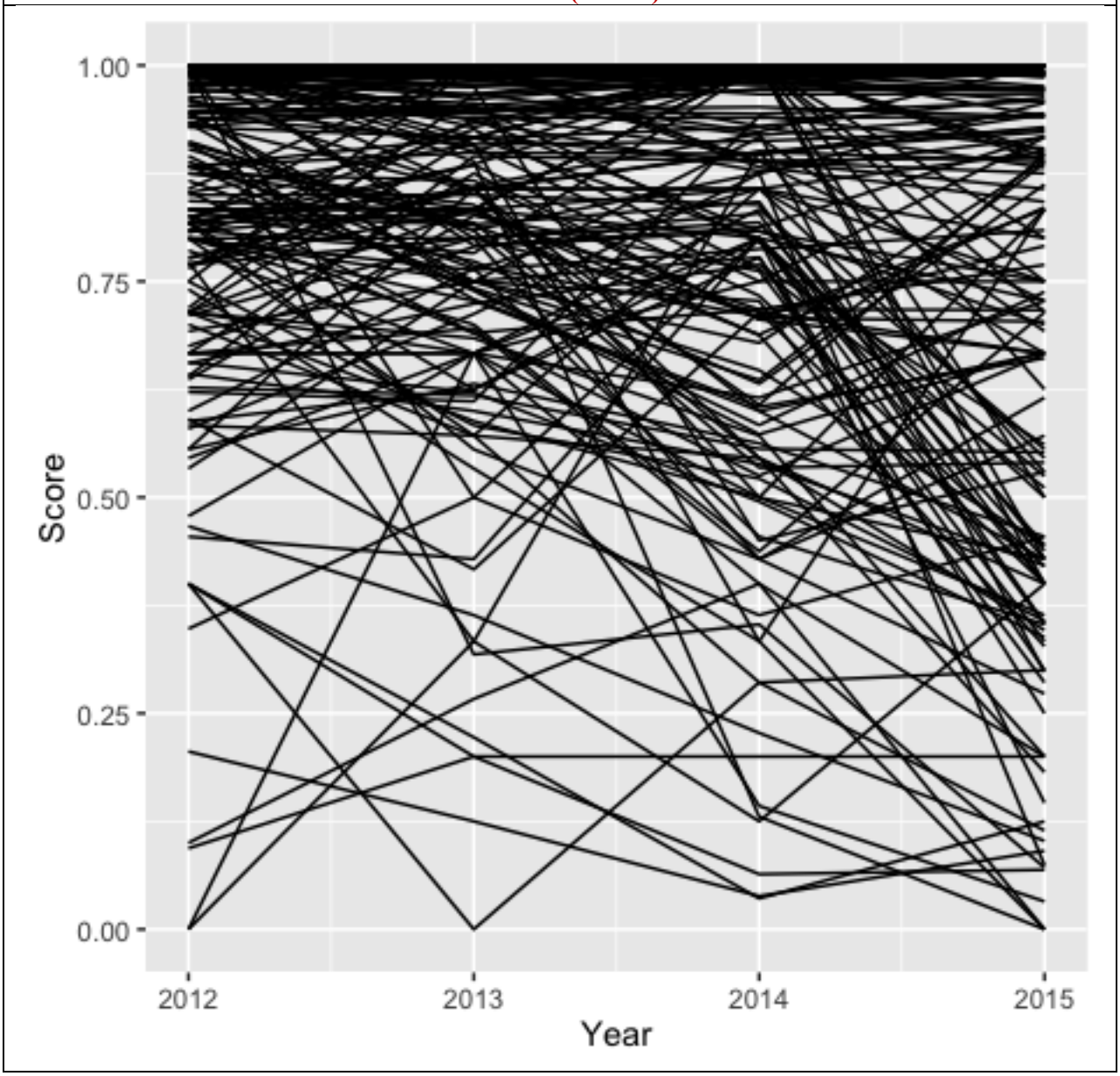
Site	Pre	Post	F	p
A	82.36	66.61	73.83	<.001
B	72.98	74.13	.147	.702
C	72.79	73.09	.005	.942
D	53.74	65.56	13.73	<.001
E	94.72	90.96	.823	.366
F	85.75	92.18	3.067	.081

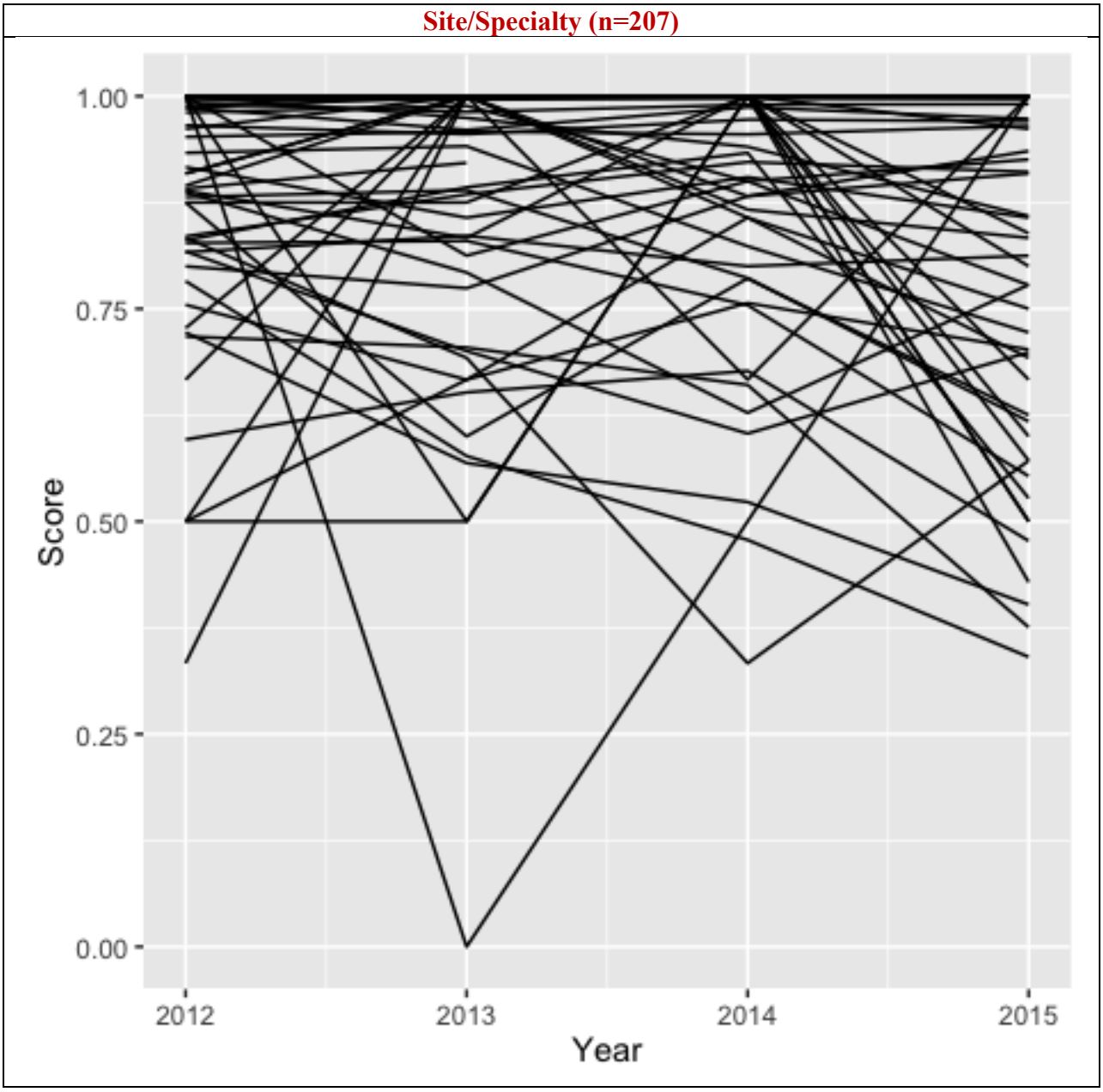
As can been seen, results were stable for the majority of sites by year and in the pre- and post-ICD-10-CM conversion period. Of note, two sites show a significant decrease over time of moderate effect size. These ongoing changes likely account for the significant variation seen across the ICD-9-CM to ICD-10-CM conversion, and antedate the conversion itself. Neither site implemented changes in EHR systems during this period expected to significanty affect feasibility of measure data elements. Further investigation into potential changes in clinical or coding practices may elucidate the reasons for variation specific to these sites, consisitent with the goal of quality measures.

Representative graphs showing year-over-year "trajectories" for providers and departments with ≥5 evaluable encounters per year are shown below, demonstrating patterns of variability in longitudinal performance.

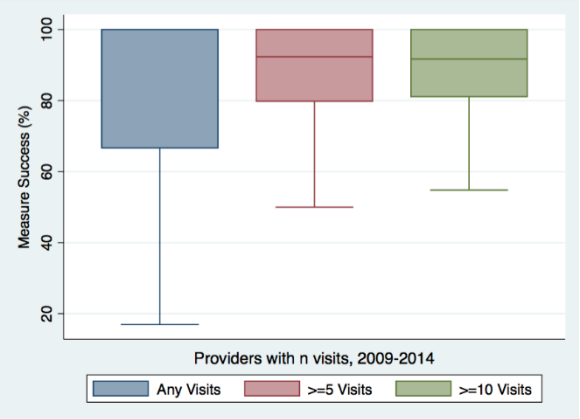
[2] Partial year (Jan 01 – Jun 30); 2 evaluable encounters required
[3] Partial year (Jan 01 – Jun 30); 2 evaluable encounters required

Provider (n=704)





The threshold of 5 visits/entity/year evaluated reveals some residual hysteresis; as expected, increases in the threshold decrease the distribution width, with the majority of the benefit accrued between 1 and 5:



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

The structure of the measure is plausibly and directly related to quality of care (*i.e.* antibiotic avoidance) as described by a clinical guideline based on extensive, high-quality evidence.

Empiric results across a large number of patients and providers at six test sites demonstrate a wide dynamic range for the measure score, indicating that it is capable of detecting differences in performance. This was observed for comparisons both within and between sites. Of note, two test sites demonstrated consistently higher scores than others; these sites also had smaller overall numbers of included encounters, which were drawn predominantly from specialty practices. Two other sites show differences over time in overall performance, as noted above. We have extensively reviewed measure computation process for these site, and do not detect errors on manual review of sampled data. As with any measure, however, an outlying value raises the question of incorrect ascertainment of measure components; while this does not compromise the overall feasibility of the measure, further investigation may provide additional insight into unintended reasons for variation in performance.

Mean scores were high, but appreciably different from ideal, and lower than would be expected solely from the prevalence of alternate indications for antibiotics not measured, indicating that current practice has opportunity in many cases for measurable improvement. Conversely, entities do reach scores >0.95, and the third quartile for several sites reaches 1.00 in data through 2016, confirming that it is possible to reach very high measure scores in practice; the effective "ceiling" set by gaps in data capture does not greatly limit the dynamic range of the measure. This is true for clinical specialties ranging from primary care, where most children with OME are seen, to otorhinolaryngology, where more complex cases are evaluated; the higher mean scores for the latter case may imply that greater experience correlates with higher measure scores, similar to results seen for all entities with higher evaluable visit counts.

2b3. EXCLUSIONS ANALYSIS

NA ☐ ☐ no exclusions — *skip to section 2b4*

The single exclusion criterion at the encounter level is described above.

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

The distribution of other diagnoses at visits meeting both inclusion and exclusion criteria, and at which antibiotics were prescribed, was examined to assess adequacy of specified exclusion diagnoses. Results were assessed qualitatively.

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

Other than those specified as exclusion criteria, no alternate diagnosis with a likely bacterial etiology was found in >4% percent of encounters in the sample undergoing records review with prescribed antibiotics. Otherwise eligible encounters that did contain ≥1 exclusion diagnoses resulted in antibiotic prescription rates of:

Site	% abx, excluded enc	% abx, all enc
A	10	13
B	35	13

C	32	23
D	53	33
E	4	2
F	11	3

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

The specified exclusion criterion accounts for the majority of alternate reasons for antibiotic prescription in otherwise eligible visits, and identifies a subset of encounters with a higher antibiotic prescription rate than other eligible encounters. Therefore, the presence of the exclusion criterion will reduce confounding of the measure score by indication.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

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

☒ **No risk adjustment or stratification**

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

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

☐ **Other,** Click here to enter description

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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

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

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

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

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

If stratified, skip to 2b4.9

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

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

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

2b4.9. Results of Risk Stratification Analysis:

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

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

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

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

See description of concurrent validity testing above.

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

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

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

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

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

one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

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

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

See feasibility and validity testing discussion and results above.

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

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

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)

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)*
ALL data elements are in defined fields in electronic health records (EHRs)

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.**

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.**
Attachment **Attachment:** 5._OME_eMeasure_Feasibility_Scorecard_FINAL.docx

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
We have developed this measure specifically to use data elements specified and populated at high frequency in electronic health records, both due to health system operational requirements and consonant with the CMS EHR Incentive ("Meaningful Use") program. Recognizing that the measure addresses a condition with moderate population frequency, we have not extended the specification to capture rare indications for antibiotic use in children; doing so would greatly increase the complexity of measure evaluation for marginal return in score improvement in most clinical contexts. Feasibility of included data elements is high in multiple health systems tested, and values are drawn from standard terminologies.

Of note with regard to visit diagnoses: For billing purposes, in most health systems ICD-9-CM was used by source systems to represent diagnoses, with conversion to use of ICD-10-CM on or about September 30, 2015. Clinician entry of diagnoses into the EHR is likely to have been recorded using an interface terminology such as Intelligent Medical Objects, rather than either of the ICD terminologies. The CMS Meaningful Use initiative requires that for data exchange users of CEHRT be able to express diagnoses in SNOMED-CT. Value sets using each of these three terminologies have been published via VSAC, to allow for more flexible evaluation of the measure.

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)*.
None

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.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement (Internal to the specific organization)	

4a.1. For each CURRENT use, checked above, provide:

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

NA – new measure

4a.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?*)

Newly developed measure

4a.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.*)

Federal and State agencies use a comprehensive set of quality measures to track the performance of the health system for children and identify areas needing improvement. CHIPRA set in motion a series of initiatives that have led to a multifaceted national effort to (a) develop valid and reliable measures of quality of care for children, (b) encourage State Medicaid and CHIP agencies to report these measures to CMS, and (c) promote the use of these measures to improve quality of care for children enrolled in Medicaid and CHIP. The CHIPRA quality demonstration projects represent examples of implementing quality measures nationwide to achieve these goals. An OME emeasure was initially included, but proved infeasible due to specifications not consistent with EHR practice, which led to the deactivation of the measure. This measure has been developed to address the same quality facet – avoidance of antibiotics to treat OME – in a manner compatible with EHR implementation. It is suitable for inclusion in the CHIPRA measure universe.

4b. 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.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Not Applicable

4b.2. 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.

Not Applicable

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Not Applicable

5. Comparison to Related or Competing Measures

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

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

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0655 : Otitis Media with Effusion: Antihistamines or decongestants – Avoidance of inappropriate use
0656 : Otitis Media with Effusion: Systemic corticosteroids – Avoidance of inappropriate use
0657 : Otitis Media with Effusion: Systemic antimicrobials – Avoidance of inappropriate use

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

5a. Harmonization
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 completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The existing NQF #0657: Otitis Media with Effusion: Systemic antimicrobials – Avoidance of inappropriate use is based on manual chart review; measure specifications are therefore not comparable, though both measures address the same aspect of clinical practice.

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

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

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Children´s Hospital of Philadelphia Pediatric Quality Measures Program Center of Excellence
Co.2 **Point of Contact:** Charles, Bailey, baileyc@email.chop.edu, 267-426-1389-
Co.3 **Measure Developer if different from Measure Steward:** The Children´s Hospital of Philadelphia Pediatric Quality Measures Program Center of Excellence
Co.4 **Point of Contact:** Charles, Bailey, baileyc@email.chop.edu, 267-426-1389-

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.
Charles Bailey, MD, PhD, The Children´s Hospital of Philadelphia (baileyc@email.chop.edu)
Role: Dr. Bailey served as the informatics lead for development and testing of the ENT measures.

Elizabeth Earley, The Children´s Hospital of Philadelphia (liz.earley@gmail.com)
Role: Ms. Earley served as the project coordinator for initial development and testing of the ENT measures.

Megan O'Karma, The Children´s Hospital of Philadelphia (okarma@email.chop.edu)
Role: Ms. O'Karma served as the project coordinator for final testing of the ENT measures.

Hanieh Razzaghi, The Children´s Hospital of Philadelphia (razzaghih@email.chop.edu)
Role: Ms. Razzaghi served as the data analyst for final testing of the ENT measures.

Brandon Becker, The Children´s Hospital of Philadelphia (beckerb1@email.chop.edu)
Role: Mr. Becker served as the biostatistician for the development of the ENT measures.

Christopher Forrest, MD, PhD The Children´s Hospital of Philadelphia (forrestc@email.chop.edu)
Role: Dr. Forrest served as a site co-principal investigator for the development of the ENT measures.

JeanHee Moon, PhD, MPH The Children´s Hospital of Philadelphia (moonj1@email.chop.edu)
Role: Dr. Moon served as the overall project manager for the Children's Hospital of Philadelphia Pediatric Quality Measures Program Center of Excellence. She also contributed directly to evaluation of the ENT measures.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

Appendix B: Additional Information Memo - #2640

Dear EENT Standing Committee,

Since we released the submission materials for measures #2640 and #2811 to you, we have received additional information from the developer that has addressed many of our concerns. Because of the short turn-around time, we are providing this information—as well as our preliminary analysis of the new information—in this "addendum". The developer will officially modify their submission materials after the March 14, 2017 webinar.

--NQF Staff

For both #2640 AND #2811:

- Now specified for only three levels of analysis: individual clinicians ["provider"], clinician practices ["department/group"], and facilities ["institution"]
- ***Still need to clarify*** whether a provider/department/institution must have more than 5 eligible encounters in the measurement time period in order to be eligible for the measure

Measure #2640 [Otitis Media with Effusion - Antibiotics Avoidance]

Reliability

*Updated **Reliability** Testing Results from the developer*

OME, antibiotic avoidance:

Entity	N	F	P
Provider	1,786	26.58	<0.0001
Department/Group*	170	124.9	<0.0001
Institution	6	2,668	<0.0001

* Because of the possibility that providers might rotate among clinics, department/group is conservatively defined as a particular special at a single institution.

NQF Preliminary Analysis

The overall method is appropriate and the updated analysis was conducted for the levels of analysis as specified.

The value 1-1/F can be considered an "average reliability". A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

Entity	**N**	**F**	**P**	**1-1/F ("average reliability")**
Provider	1,786	26.58	<0.0001	0.9624

Department/Group*	170	124.9	<0.0001	0.9920
Institution	6	2,668	<0.0001	0.9996

Guidance from the Reliability Algorithm

Specifications are precise (Box 1) → Empirical testing conducted for all three levels of analysis specified (Box 2) → Score-level testing was conducted (Box 4) → Method is appropriate (Box 5) High certainty that the performance measure scores are reliable (Box 5a) → High

The highest possible rating is HIGH.

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

Validity

Updated Validity Information from the developer

Data element validity testing: Further information on the analysis of 225 encounters from one site: The data from this site includes 28 primary care practices (4 hospital-based and 24 community-based) with largely practice-specific staff; 21 specialty departments were also included in the OME dataset.

Score-level testing: From the score reliability/discriminant ability testing, we know that sites are different groups from the measure's perspective; this has face validity as well, since we expect that differences in practice and training across institutions will underlie the differences in measure results.

Hypothesis: the same providers will, absent external influences, practice in a consistent way over time.

Rationale for hypothesis:

- The consensus best practice did not change over the interval we're examining (i.e., the specialty society guidelines we're tracking have not changed in respects important to the measure over the interval)
- The technical infrastructure hasn't changed qualitatively (there've certainly been upgrades and optimizations to the EHRs, and to hospital infrastructure, but none at the level of, say, changing EHRs)
- No evidence that there have been attempts to change practice by institutions or payers.

Expected results based on this hypothesis: Measure scores within a given provider (and consequentially for groups of providers) will vary less over time than scores between different providers or groups. In terms of the ANOVA, the hypothesis predicts that the year-over-year F statistic will be smaller than the group-to-group F statistic.

Results of testing:

Entity	F (between entity)	F (between year)
Provider (threshold = 5)	151.3	14.5
Provider (threshold = 10)	21.1	1.9
Department	124.9	1.9

Conclusion: Hypothesis is supported.

NQF Preliminary Analysis

Developers hypothesized that measure results within individual providers/departments would vary less across time than measure results between providers/departments, given the lack of external influences that would affect results across time. This can be considered a form of score-level construct validation.

The analogous analysis for the facility [institution/site] level of analysis was not provided. It is not clear if/how the results of the year-by-site testing analysis that was initially presented support the stated hypothesis. As noted in the initial staff preliminary analysis, that analysis itself would be considered a weak form of construct validation (i.e., comparing the score with itself across time) if the data were not aggregated by provider/department and then by site.

Guidance from the Validity Algorithm

Specifications are consistent with the evidence (Box 1) → Potential threats to validity were assessed (Box 2) → Empirical testing was conducted for at two of the three levels of analysis specified (Box 3) → Score-level testing was conducted for at least two of the three levels of analysis specified [facility-level testing results not clear] (Box 6) → Method is described and seems appropriate (Box 7) → Moderate certainty that measure scores are a valid indicator of quality (Box 8b) → Moderate [ASSUMING site-level results can be explained by the developer]

NOTE: The chart review analysis and the correlation analysis comparing measure results using ICD-9-CM coding versus using a proprietary coding system support the validity of the measure but cannot stand alone because the data were derived from only one EHR (NQF requires testing from more than one EHR).

The highest possible rating is HIGH.

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

Appendix C: Details of Measure Evaluation - #2640

2640: Otitis Media with Effusion - Antibiotics Avoidance

Submission

Description: The proportion of encounters with a diagnosis of Otitis Media with Effusion (OME) made at age 2 months to 12 years, where patients were not prescribed systemic antimicrobials.

Numerator Statement: Eligible encounters at which a systemic antibiotic was not prescribed.

Denominator Statement: Outpatient encounters at which otitis media with effusion is diagnosed, but at which common conditions for which antibiotics are indicated are not diagnosed. It is expected that a small fraction of patients with rare non-OME indications for antibiotic usage will not be identified by the specified exclusion criteria, but these will be rare cases, and will not alter the measure score significantly in most practice contexts. Of note, however, applicability may be limited in specific practice environments in which a large proportion of patients seen have immune deficiencies requiring chronic antibiotic use (e.g. immunology or hematology/oncology clinics).

Exclusions: Diagnosis at the visit of common childhood infection for which antibiotics are frequently indicated.

Adjustment/Stratification: No risk adjustment or risk stratification Measure validity is not dependent on stratification, but an organization may consider stratifying by sociodemographic factors in order to assess disparities in care provide in Otitis Media with Effusion.

Level of Analysis: Facility, Clinician : Group/Practice, Clinician : Individual, Integrated Delivery System

Setting of Care: Clinician Office/Clinic, Urgent Care - Ambulatory

Type of Measure: Process

Data Source: Other

Measure Steward: The Children's Hospital of Philadelphia Pediatric Quality Measures Program Center of Excellence

STEERING COMMITTEE MEETING [03/14/2017]

1. Importance to Measure and Report: <u>The measure meets the Importance criteria</u>

(1a. Evidence, 1b. Performance Gap)

1a. Evidence: H-10; M-0; L-0; I-0; 1b. Performance Gap: H-0; M-6; L-4; I-0**;**

<u>Rationale</u>:

- The developer provided a clinical practice guideline recommendation against using systemic antibiotics for treating Otitis Media with Effusion (OME). The recommendation, graded as “strong” and supported by grade A evidence, is published in two peer-reviewed publications: The Pediatrics Journal (2004) and Otolaryngology Head and Neck Surgery (2016).
- Based on data from 36,060 visits documented in the Children´s Hospital of Philadelphia’s electronic health record from 2009-2014, the provider-level “mean failure rate” reported by the developer was 15.05% and the facility-level rate was 11.42%. ○ Committee members questioned the meaning of the 15% mean failure rate. The developer clarified that for providers the average provider-level performance rate for the measure is approximately 85%, meaning that, on average, providers prescribed an antibiotic 15% of the time when the patient had a diagnosis of OME but no other conditions that might require antibiotics.
    - The Committee noted that the performance rate (85%) was relatively high, and questioned the ability to improve performance. The developer noted that approximately 25% of providers included in their testing data are achieving 100%, suggesting it is possible for other providers to do so.
- Members questioned whether there were any differences in performance for particular population subgroups (e.g., ethnicity, race, sex, social economic status). The developers reported finding relatively

small, but statistically significant, differences in provider-level performance between racial/ethnic groups and those with varying insurance status/type. However, they did not provide the data from these analyses.

- Several Committee questioned the need for this measure, noting the decrease in the incidence of OME over the past several years. However, the developer noted that otitis media is "the primary driver of antibiotic prescriptions" in their dataset.

2. Scientific Acceptability of Measure Properties: <u>The measure [does not] meet the Scientific Acceptability criteria</u>

(2a. Reliability - precise specifications, testing; 2b. Validity - testing, threats to validity)

2a. Reliability: H-0; M-2; L-8; I-0 2b. Validity: **H-X; M-X; L-X; I-X**

<u>Rationale</u>:

- The developer conducted score-level reliability testing for the three levels of analysis specified for the measure. Results indicate an average reliability of >0.96 for all three levels of analysis. These results are based on data for January 2009-June 2016 from 6 academic pediatric health systems, 704 clinicians and 207 practices, and 3 EHR systems.
- Committee members had a lengthy discussion regarding the difficulty in accurately diagnosing of Otitis Media with Effusion. Several members expressed concern regarding the potential of "gaming the system" by inappropriately coding as Acute Otitis Media (rather than OME) if they have decided to prescribe antibiotics. The developer agreed that accurate diagnosis is a problem, but pointed out that the measure is designed to assess prescription of antibiotics when the clinician has diagnosed as OME.
- After much discussion, the Committee agreed that the measure did not pass the reliability subcriterion and did not recommend the measure for endorsement.

STAFF NOTE: The discussion of Committee regarding ability to diagnosis OME is more properly one of validity rather than reliability. The developer provided some information relevant to this discussion under the validity subcriterion. NQF will ask the Committee to re-vote on reliability, basing its rating on clarity of specifications and results of reliability testing, and to consider the question of diagnosis accuracy and the data provided by the developer in a discussion on validity.

3. Feasibility: H-X; M-X; L-X; I-X

(3a. Clinical data generated during care delivery; 3b. Electronic sources; 3c.Susceptibility to inaccuracies/ unintended consequences identified 3d. Data collection strategy can be implemented)

<u>Rationale</u>:

-

4. Usability and Use: H-X; M-X; L-X; I-X

(Used and useful to the intended audiences for 4a. Accountability and Transparency; 4b. Improvement; and 4c. Benefits outweigh evidence of unintended consequences)

<u>Rationale</u>:

-

5. Related and Competing Measures

-

Steering Committee Recommendation for Endorsement: **Not Recommended**

<u>Rationale:</u>

-

2640: Otitis Media with Effusion - Antibiotics Avoidance
6. Public and Member Comment •
7. Consensus Standards Approval Committee (CSAC) Vote: Y-X; N-X
8. Appeals