

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Otitis Media with Effusion Systemic Antimicrobials Avoidance of Inappropriate Use

Date of Submission: 3/27/2015

Type of Measure:

| | |
|---|---|
| <input type="checkbox"/> Composite – STOP – use composite testing form | <input type="checkbox"/> Outcome (including PRO-PM) |
| <input type="checkbox"/> Cost/resource | <input checked="" type="checkbox"/> Process |
| <input type="checkbox"/> Efficiency | <input type="checkbox"/> 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 specifications** (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 (*including questions/instructions*; minimum font size 11 pt; do not change margins). **Contact NQF staff if more pages are needed.**
- Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](#).

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

2a2. Reliability testing ¹⁰ demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b2. Validity testing ¹¹ demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

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

AND

If patient preference (e.g., informed 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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; ^{14,15} and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

Notes

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

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

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

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

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

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

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

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

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

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

| Measure Specified to Use Data From: (must be consistent with data sources entered in S.23) | Measure Tested with Data From: |
|---|---|
| <input checked="" type="checkbox"/> abstracted from paper record | <input checked="" type="checkbox"/> abstracted from paper record |
| <input type="checkbox"/> administrative claims | <input type="checkbox"/> administrative claims |
| <input type="checkbox"/> clinical database/registry | <input type="checkbox"/> clinical database/registry |
| <input type="checkbox"/> abstracted from electronic health record | <input type="checkbox"/> abstracted from electronic health record |
| <input type="checkbox"/> eMeasure (HQMF) implemented in EHRs | <input type="checkbox"/> eMeasure (HQMF) implemented in EHRs |
| <input type="checkbox"/> other: Click here to describe | <input type="checkbox"/> other: Click here to describe |

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

Not applicable

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

AMA-PCPI Testing Project

- Data collected from patients seen between February 1, 2008 and January 31, 2009

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

| Measure Specified to Measure Performance of: (must be consistent with levels entered in item S.26) | Measure Tested at Level of: |
|---|--|
| <input checked="" type="checkbox"/> individual clinician | <input checked="" type="checkbox"/> individual clinician |
| <input checked="" type="checkbox"/> group/practice | <input checked="" type="checkbox"/> group/practice |
| <input type="checkbox"/> hospital/facility/agency | <input type="checkbox"/> hospital/facility/agency |
| <input type="checkbox"/> health plan | <input type="checkbox"/> health plan |
| <input type="checkbox"/> other: Click here to describe | <input type="checkbox"/> other: Click here to describe |

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

The data sample came from two pediatric practice network groups. The first network represented a range of practice settings, sizes, locations and medical record systems in the United States. The second network had approximately 151 pediatrician members from 41 states in diverse practice settings and 65 community child health care providers in 30 practices.

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

AMA-PCPI Testing Project

- A reviewer was instructed to identify eligible medical records for 30 patients with OME and then a second reviewer independently re-abstracted 10 of the records for tests of inter-rater reliability (IRR)
- 19 practice sites contributed abstraction data, representing urban, suburban, and rural locations
 - 10 sites used a paper record system
- A total of 114 re-abstractions for inter-rater reliability were performed during the project
- Data abstraction performed between March 1, 2009 and June 30, 2009

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

The data sample used for reliability testing was acquired from two pediatric practice network groups.

The data sample used for validity testing was acquired through survey of an expert panel consisting of 21 providers from varying specialties.

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)*

Data abstracted from randomly sampled patient records were used to calculate inter-rater reliability for the measure. Charts for abstraction were selected for OME patients aged 2 months to 12 years who had at least 1 visit between February 1, 2008 and January 31, 2009.

Data analysis included:

- Percent agreement
- Kappa statistic to adjust for chance agreement

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

Reliability: N, % Agreement, Kappa (95% CI)

Numerator: 62, 95%, 0.00 (0.00, 0.00)*

Denominator: 115, 74%, 0.41 (0.24, 0.59)

Exceptions: N/A, N/A, N/A, N/A

Overall: 62, 95%, 0.00 (0.00, 0.00)*

*This is an example of the limitation of the Kappa statistic. While the agreement can be 90% or greater, if one classification category dominates, the kappa can be significantly reduced.

(<http://www.ajronline.org/cgi/content/full/184/5/1391>)

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?)

Although the kappa for this measure seems low, after further investigation, the low kappa was found to result from the limitation of the Kappa statistic where while the agreement can be 90% or greater, if one classification category dominates, the kappa can be significantly reduced.

(<http://www.ajronline.org/cgi/content/full/184/5/1391>)

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 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)

Face validity of the measure score as an indicator of quality was systematically assessed as follows. After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 5= Strongly Agree

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

Our expert panel included 21 members, including representatives of the AAO-HNS Patient Safety and Quality Improvement (PSQI) committee, Performance Measure Task Force (PMTF), and members of the AOE and OME guideline development panels. The list of expert panel members is as follows:

Emily Boss, MD

Rahul K. Shah, MD, MBA
David Roberson, MD
Carl Snyderman, MD, MBA
Spencer Payne, MD
Seth Schwartz, MD, MPH
Julie L. Goldman, MD
C. Ron Cannon, MD
Michael McCormick, MD
Jesse Hackell, MD
David Hoelting, MD
Maria Veling, MD
Jonathan Kopelovich, MD
Geoffrey Simon, MD
Richard V. Smith, MD
Dennis Poe, MD
Robert Stachler, MD
K. Ashok Kumar, MD, FRCS, FAAFP
Lisa Hunter, PhD
Giri Venkatraman, MD
Ryan Sewell, MD, JD

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 results of the expert panel rating of the validity statement were as follows: N = 21; Mean rating = 4.43 and 90.1% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings
1 – 0 responses (Strongly Disagree)
2 – 1 response
3 – 1 response (Neither Agree nor Disagree)
4 – 7 responses
5 – 12 responses (Strongly Agree)

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — skip to section 2b4

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*)

- Analysis of exceptions were not included as part of the inter-rater reliability testing project.

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*)

- Analysis of exceptions were not included as part of the inter-rater reliability testing project.

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*)

- Analysis of exceptions were not included as part of the inter-rater reliability testing project.

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.

[Not applicable](#)

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of $p < 0.10$; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

[Not applicable](#)

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

[Not applicable](#)

2b4.4b. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps – do not just name a method; what statistical analysis was used*)

[Not applicable](#)

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](#)

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

2b4.9. Results of Risk Stratification Analysis:

Not applicable

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

Not applicable

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

Not applicable

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

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

- This measure was not tested for meaningful differences in performance across providers or practice sites

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*)

- This measure was not tested for meaningful differences in performance across providers or practice sites

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?*)

- This measure was not tested for meaningful differences in performance across providers or practice sites

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

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

Note: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

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

- Not applicable

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

- Not applicable

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

- Not applicable

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 non-responders) and how the specified handling of missing data minimizes bias (describe the steps – do not just name a method; what statistical analysis was used)

This measure was found to be reliable for implementation through paper-based modalities. All key data elements for the measure were identified during reliability testing.

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)

Not Applicable

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 non-responders) 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)

Not Applicable