

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

Measure Title: Selection of Prophylactic Antibiotic—First OR Second Generation Cephalosporin

Date of Submission: March 17, 2014

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 type="checkbox"/> abstracted from paper record	<input type="checkbox"/> abstracted from paper record
<input checked="" type="checkbox"/> administrative claims	<input checked="" 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).

Physician Quality Reporting System (PQRS) 2010 data.

1.3. What are the dates of the data used in testing? Click here to enter date range

The dates for the data used in testing were collected from patients seen in 2010.

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)

Data used for this sample originated from an alliance of 120 academic medical centers and over 250 affiliated hospitals located in an urban setting in a large Midwestern city. The number of physicians per site ranged from 400-1000 physicians.

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)

The total number of physicians reporting on this measure used for this analysis was 2,125. Of those, 306 met the minimum number quality reporting events for inclusion in the reliability analysis.

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.

Not applicable.

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) Almost always pick 1 unless its signal to noise where we pick 2

☒ **Performance measure score** (e.g., signal-to-noise ratio 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)

Signal-to-noise ratio analysis

Reliability testing of the Perioperative measures claims was carried out using an application of the signal-to-noise ratio (SNR) measure. In this analysis, reliability equals the ratio of signal to noise, at the physician level. The signal is the variability in measured performance that can be explained by real differences in physician performance; noise is the total variability in measured performance. Reliability is then the ratio of the physician-to-physician variance to the sum of the physician-to-physician variance plus the error variance specific to a physician:

$$\text{Reliability} = \text{Variance (physician-to-physician)} / [\text{Variance (physician-to-physician)} + \text{Variance (physician-specific-error)}]$$

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician's true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

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)

Signal-to-noise ratio analysis

For this measure, the reliability at the minimum level of quality reporting events was 0.8875. The reliability at the average number of quality reporting events was 0.9677.

This measure has high reliability when evaluated at the minimum level of quality reporting events and at the average number of quality events.

Table 1.

Measure	Reliability at Minimum Number of Events, N= 10	Reliability at Average Numbers of Events	Numbers of Physicians	Performance
Selection of Prophylactic Antibiotic- First OR Second Generation Cephalosporin (PQRS #21; NQF #0268)	88.75%	96.77%	306	76.13%

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?)

Signal-to-noise ratio analysis

Reliability estimates at the minimum number of quality reporting events (N=10) and at the mean number of quality reporting events per physician are reported in Table 1 above. Reliability across the measures is no less than 69% at the minimum number of events, and 91% or higher at the average number of events. The average performance rate among physicians with at least 10 quality reporting events ranges from 7.2% to 97.8% across the measures. While a performance rate of 90%+ might suggest there is limited opportunity for improvement, the rates for these physicians and the practice settings may not capture the variation in performance across other settings. As a result, monitoring performance rates should continue in these and in other practice settings.

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

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

Face Validity Perioperative Measure Expert Panel

The following representatives from American Academy of Orthopedic Surgeons, American Society of Anesthesiologists, American College of Surgeons, Society for Vascular Surgery, American College of Surgeons, American Academy of Otolaryngology- Head and Neck Surgery, American Society of Plastic Surgeons, American Association of Neurological Surgeons Congress of Neurological Surgeons, American College of Obstetrics and Gynecologists, Society of Thoracic Surgeons, and American Society of Breast Surgeons assessed the face validity of this measure:

Aamir Siddiqui, MD
Gregory Surfield, MD
Karol Gutowski, MD, FACS
Daniel T. Ness, MD, FACS
Jeffrey Landercasper MD, FACS
Robert Buras, MD
Fred H. Edwards, MD
Barbara S. Levy, MD
John Ratliff, MD
Debra Johnson, MD
Scot Glasberg, MD
Gary R. Culbertson MD, FACS
William A Wooden, MD
Lee D. Eisenberg, MD, FACS
John O. Gage, MD
Robert M. Zwolak, MD
Eric B. Whitacre, MD
Mark Savarise, MD, FACS
Peggy G. Duke, MD
William O. Shaffer MD
David Nielsen, MD

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

The results of the expert panel rating of the validity statement were as follows: N = 21, Mean rating = 4.05, and 76.2% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 (Strongly Disagree) – 1 response
2 (Disagree) – 2
3 (Neither Agree nor Disagree) –1 response
4 (Agree) – 8 responses
5 (Strongly Agree) – 9 responses

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?*)

Face validity has been quantitatively assessed for this measure. Specifically, the Perioperative Measure Expert Group members were asked to empirically assess face validity of the measure. The expert panel consisted of 21 members, whose specialties include anesthesiology, general, orthopedic, vascular, thoracic surgery, otolaryngology, obstetrics and gynecology, methodology, and plastic surgery.

76.2% of respondents, with an average score of 4.5, either strongly agree or agree that this measure can accurately distinguish good and poor quality.

Response Key

1 (Strongly Disagree)

2 (Disagree)

3 (Neither Agree nor Disagree)

4 (Agree)

5 (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*)

Signal-to-noise ratio analysis

Exceptions included documentation of medical reason(s) for not ordering a first OR second generation cephalosporin for antimicrobial prophylaxis. Exceptions were analyzed for frequency and variability across providers.

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*)

Signal-to-noise ratio analysis

There were 13,646 total quality reporting events for all physicians. Of these, 894 were excluded from consideration for physician performance because they were submitted with a modifier and met the exception criteria. This brought the total number of eligible quality reporting events to 12,752. These exclusions represented 6.55% of the total quality reporting events for all physicians.

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*)

Exceptions are necessary to account for those situations when it is not medically appropriate for a patient to receive a first or second generation cephalosporin, such as when patient has an allergy. Excluded patients account for a relatively small number of the total quality reporting events and prevent unnecessary and/or inappropriate prescription of cephalosporins.

The PCPI methodology uses three categories of reasons for which a patient may be excluded from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples have been provided in the measure exception language of instances that would constitute an exception and are intended to serve as a guide to clinicians. Rather than specifying an exhaustive list of explicit medical, patient, and system reasons for exception for each measure, the PCPI rely on clinicians to link the exception with a specific reason for the decision to not perform the action/service/intervention required by the measure. Where examples of exceptions are included in the measure language, the PCPI has specified these reasons within the measure specifications, however this list is not intended to be an exhaustive list of reasons. Some have indicated concerns with exception reporting --the potential for physicians to inappropriately exclude patients to enhance their performance statistics. Research has indicated that levels of exception reporting occur infrequently and are generally valid. (Doran et al., 2008), (Kmetik et al., 2011) Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by physicians, and it ameliorates perverse incentives to refuse care to "difficult" patients." (Doran et al., 2008)

Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommend that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. We also advocate for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

References:

Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of pay for performance targets by English Physicians. *New Engl J Med.* 2008; 359: 274-84.
Kmetik KS, Otoole MF, Bossley H et al. Exceptions to Outpatient Quality Measures for Coronary Artery Disease in Electronic Health Records. *Ann Intern Med.* 2011;154:227-234.

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. Not an outcome/resource use measure.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of $p < 0.10$; correlation

of x or higher; patient factors should be present at the start of care and not related to disparities)

Not Applicable

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

Not Applicable

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

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)

We calculated interquartile range and percentiles to demonstrate statistically significant differences in performance.

CMS 2010 Physician Quality Reporting Initiative:

Eligible Professionals: 69,130 eligible professionals
Professionals Reporting: >=1 valid QDC: 6175 professionals
% Professionals Reporting who reported >=1 valid QDC: 8.9%
Professionals Satisfactorily Reporting: 3415
% Professionals Satisfactorily Reporting: 55.3%

CMS 2008 Physician Quality Reporting Initiative:

170,155 patients were reported on for the 2008 program.
Eligible Professionals: 73,193
Professionals Reporting >=1 Valid QDC: 5929
% Professionals Reporting >=1 Valid QDC: 8.10%
Professionals Satisfactorily Reporting: 2955
% Professionals Satisfactorily Reporting: 49.84%

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)

¹Confidential CMS PQRI 2010 Performance Information by Measure. Jan-Sept TAP file.

2010

Scores on this measure: N = 69,130 Eligible Professionals; Mean = 93.53% is the mean performance rate of NPI's/Tax Identification Number.

10th percentile: 85.29%
25th percentile: 100%
50th percentile: 100%
75th percentile: 100%
90th percentile: 100%

The inter-quartile range (IQR) provides a measure of the dispersion of performance and the IQR for this measure is 0%. This indicates that 50% of physicians' performance was 100%.

²Confidential CMS PQRI 2008 Performance Information by Measure. Jan-Sept TAP file.

2008

Scores on this measure: N = 73,193 Eligible Professionals
10th percentile: 9.09%
25th percentile: 33.33%
50th percentile: 68.69%
75th percentile: 91.67%
90th percentile: 100%

The inter-quartile range (IQR) provides a measure of the dispersion of performance and the IQR for this measure is 58.34%. This indicates that 50% of physicians' performance ranged from 33.33% to 91.67%.

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?*)

PQRS data from 2010 indicates 0% interquartile range. Despite this small range, it is important to keep in mind that PQRS data only contains data for the Medicare population, though this measure captures patients as young as 18 years old. Further, PQRS data is not reliably representative of national performance, as it is based on voluntary reporting with about 29% of eligible professionals participating using any reporting option in 2011.

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