**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** 0514

**Measure Title**: MRI Lumbar Spine for Low Back Pain

**Date of Submission**: 03/03/2014 (2014 Submission) | 11/03/2016 (2016 Submission)

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process (2016 Submission) | Efficiency (2014 Submission) |
| 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](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * 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**](#Note10) 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**](#Note11) 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**](#Note12)  **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**](#Note13)  **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**](#Note14)**,**[**15**](#Note15) 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**](#Note16) **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: Click here to describe | 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*).

**2014 Submission:**

This measure was originally constructed using the 100% Medicare Fee-For-Service (FFS) Outpatient Standard Analytical Files (SAFs) from 2007. Medicare FFS SAFs were used for all subsequent calculations, with the Medicare FFS SAF from 2011 being the most recent data source.

**2016 Submission:**

We tested the measure using 2010-2013 Medicare fee-for-service (FFS) data from the 100% samples of the Outpatient Standard Analytic File (SAF-O), Inpatient Standard Analytic File (SAF-I), and Carrier File.

*Facility Analysis*

1. Datasets used to define the initial patient population (denominator):

* *SAF-O:* CORE and Lewin defined the initial patient population based on the 2013 100% SAF-O file. The initial patient population includes all claims for an MRI lumbar-spine study with a diagnosis of low back pain from January 1, 2013-December 31, 2013, provided in a hospital outpatient setting. This dataset also includes unique patient and facility identifiers.
* *Enrollment database and denominator files:* This dataset contains Medicare FFS enrollment, demographic, and death information for patients identified in the above file.
* *Provider of services (POS) file:* The POS file contains data on facility characteristics including urbanicity, bed count, and teaching status.

1. Datasets used to capture the numerator:

* *SAF-O and Carrier:* For patients included in the initial patient population, CORE and Lewin identified numerator exception cases by searching the 2012 and 2013 100% SAF-O and Carrier files for one or more claims for antecedent conservative therapy in the 60 days preceding the MRI lumbar-spine study.

1. Datasets used to identify measure exclusions:

* *SAF-O, SAF-I, and Carrier:* For patients included in the initial patient population, CORE and Lewin identified denominator exclusions by searching the 2010-2013 100% SAF-O, SAF-I, and Carrier files for risk factor diagnoses in the three years preceding the MRI lumbar-spine study.

**1.3. What are the dates of the data used in testing**? 2007-2011(2014 Submission) | January 2010 – December 2013 (2016 Submission)

**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: state, national | other: state, national |

**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*)

**2014 Submission:**

Measure percentages were calculated for all hospital facilities in the Medicare FFS SAF files. Inclusion of facilities in subsequent analyses varied by type and intent of analysis. Analyses describing publicly reported values included all facilities eligible in the Hospital Outpatient Quality Reporting (HOQR) Program, regardless of whether the facility chooses to participate in the program or not. There are a total of 3,680 eligible facilities in the HOQR Program, which include short-term, acute care hospitals, as well as critical access hospitals (CAHs). Case count requirements were applied to exclude those facilities that did not have a significant number of cases for this measure for some analyses.

**2016 Submission:**

The testing sample included 2,569 facilities. The number of measured entities (facilities) varies by testing type; see **Section 1.7** for details.

**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*)   
**2014 Submission:**

All Medicare FFS patients in the SAF file were included in the testing and analysis. Some analyses only included facilities and patients eligible for the Hospital Outpatient Quality Reporting (HOQR) Program.

**2016 Submission:**

The number of patients varies by testing type; see **Section 1.7** for details. Prior to applying minimum case count, there were 521,460 MRI lumbar spine study denominator cases in the hospital outpatient (facility) setting.

**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**.

**2014 Submission:**

Measure testing relied exclusively on 100% Medicare FFS SAF data. The data year and subset used for each level of testing is reported in the testing narrative for each section.

**2016 Submission:**

The data sources, dates, number of measured entities, number of MRI lumbar-spine studies, number of antecedent conservative therapies, level of analysis, and demographic profile for the patients used in each type of testing are as follows:

**Reliability Testing**

*Facilities*

Data Source: Denominator: SAF-O, SAF-I, and Carrier; Numerator: SAF-O and Carrier; Exclusions: SAF-O, SAF-I, and Carrier

Dates: Denominator: January 1, 2013-December 31, 2013; Numerator: November 1, 2012-December 31, 2013; Exclusions: January 1, 2010-December 31, 2013

Number of Measured Entities: 1,616

Number of MRI Lumbar-Spine Studies: 164,848

Number of Antecedent Conservative Therapy Cases: 62,009

Level of Analysis: Facility

Patient Characteristics: Gender (% Male): 42.4; Mean Age (Years): 66.5 (St. Dev.: 12.2); Race/Ethnicity (% Minority): 14.4

**Validity Testing**

Data Source: Structured qualitative survey questions completed by technical expert panel (TEP) members regarding measure face validity

Dates: June-July 2015

Number of Responses: 11

Respondent Characteristics: CORE and Lewin asked respondents to select at least one of the following categories: insurer/purchaser (3); payer (1); clinician (6); management/administration (5); patient/patient advocate/caregiver (3).

**Exclusions Analysis**

*Facilities*

Data Source: Denominator: SAF-O, SAF-I, and Carrier; Numerator: SAF-O and Carrier; Exclusions: SAF-O, SAF-I, and Carrier

Dates: Denominator: January 1, 2013-December 31, 2013; Numerator: November 1, 2012-December 31, 2013; Exclusions: January 1, 2010-December 31, 2013

Number of Measured Entities: 2,569

Number of MRI Lumbar-Spine Studies: 521,460

Number of Antecedent Conservative Therapy Cases: 356,163

Level of Analysis: Facility

Patient Characteristics: Gender (% Male): 39.3; Mean Age (Years): 68.3 (St. Dev.: 12.3); Race/Ethnicity (% Minority): 12.7

**Risk Adjustment/Stratification**

N/A; this measure is not risk adjusted or risk stratified.

**Identification of Statistically Significant & Meaningful Differences in Performance**

*Facilities*

Data Source: Denominator: SAF-O, SAF-I, and Carrier; Numerator: SAF-O and Carrier; Exclusions: SAF-O, SAF-I, and Carrier

Dates: Denominator: January 1, 2013-December 31, 2013; Numerator: November 1, 2012-December 31, 2013; Exclusions: January 1, 2010-December 31, 2013

Number of Measured Entities: 1,616

Number of MRI Lumbar-Spine Studies: 164,848

Number of Antecedent Conservative Therapy Cases: 62,009

Level of Analysis: Facility

Patient Characteristics: Gender (% Male): 42.4; Mean Age (Years): 66.5 (St. Dev.: 12.2); Race/Ethnicity (% Minority): 14.4

**Comparability of Performance Scores when More than One Set of Specifications**

N/A; this measure only relies on one set of specifications.

**Missing Data Analysis and Minimizing Bias**

N/A; this measure is calculated using cleaned, post-adjudicated claims data for which no missing data was observed.

**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).*

We assessed patient-level SDS factors as part of the regression model reported in **Section 1b.4**, which provides an overview of disparities in care for patient sub-populations. CORE and Lewin based this analysis on SDS variables included in the CMS Patient Eligibility file:

* Age group
* Gender
* Race

**2016 Submission:**

While an analysis of SDS factors is important in understanding differences in care for patient sub-populations, this measure is a process measure that is neither risk adjusted nor risk stratified. Risk adjustment or risk stratification would not be appropriate based on the measure evidence base and the measure construct. Additional information on this determination is provided in **Section 2b4.2**.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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*)

**2014 Submission:**

Reliability was calculated in accordance with the methods discussed in *The Reliability of Provider Profiling: A Tutorial* (2009). The reliability testing calculates the ability of the measure to distinguish between the performance of different facilities. Specifically, the testing calculated the signal-to-noise ratio for each facility meeting the minimum case count and other criteria for public reporting from the 2011 FFS SAF file. The reliability score is estimated using a beta-binomial model, which is appropriate for the reliability testing of pass/fail measures. The reliability score for each facility is a function of the facility’s sample size and score on the measure, and the variance across facilities.

*Reference:*

Adams JL. The reliability of provider profiling: a tutorial. Santa Monica, CA: RAND Corporation. 2009. Retrieved from http://www.rand.org/pubs/technical\_reports/TR653.

**2016 Submission:**

We calculate reliability in a manner consistent with NQF guidance and in accordance with the methods discussed in *The Reliability of Provider Profiling: A Tutorial* (2009). The reliability testing calculates the ability of the measure to distinguish between the performances of different facilities. Specifically, the testing calculated the signal-to-noise ratio for each facility meeting the minimum case count in 2013. CORE and Lewin estimate the reliability score using a beta-binomial model, which is appropriate for the reliability testing of dichotomous measures (where numerator cases may only have a value of zero or one). The reliability score for each facility is a function of the facility sample size and score on the measure, and the variance across facilities.

**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*)

**2014 Submission:**

*Figure 1* (below) is a histogram of the distribution of the reliability scores for the facilities meeting all public reporting requirements in 2011. Reliability scores ranged from 24.8% to 91.8%, with a median reliability score of 53.1%.

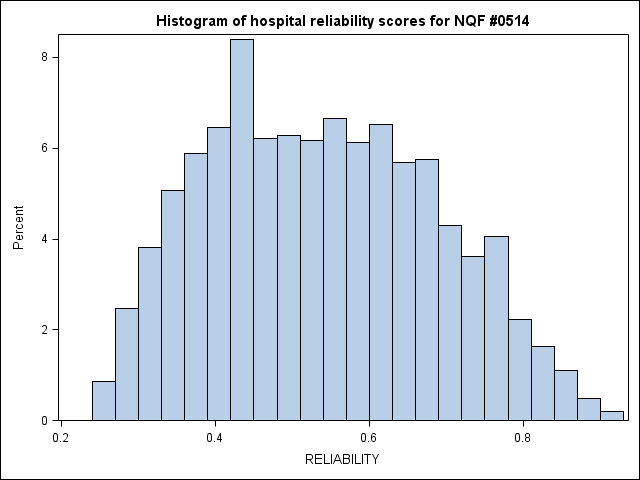


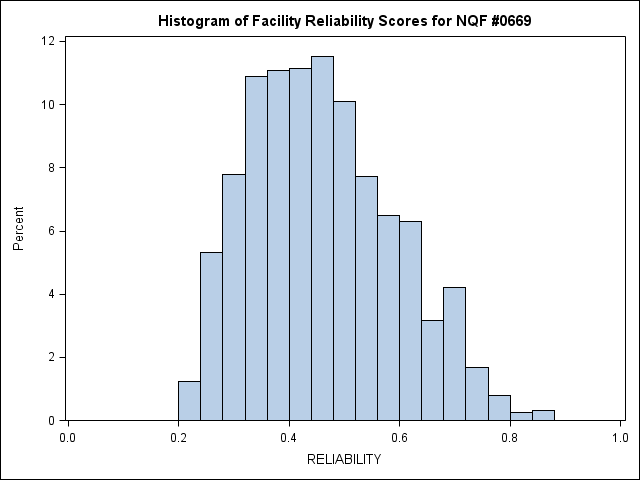
Figure 1: Histogram of Reliability Scores

**2016 Submission:**

***Facility Reliability***

*Figure 1* (below) is a histogram of the distribution of the reliability scores for the facilities meeting the minimum case count in 2013. Reliability scores ranged from 22.4% to 86.6%, with a median reliability score of 44.9%.

*Figure 1*



**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?*)  
**2014 Submission:**

A median reliability score of 53.1% is below the target median value using a beta-binomial model if the intent of the measure is to identify differences between individual facilities. However, the intent of the measure is not to identify differences in performance between individual facilities, but, rather, to identify differences from the mean (or threshold) performance value. Thus, the measure contractor purports that the testing reported in **Section 2b5**, to determine statistically significant and meaningful differences in performance, is a more appropriate test for this measure.

**2016 Submission:**

The results of reliability testing are similar to the results reported to NQF in 2014. During the August 2014 review of the measure, the working group classified the measure’s reliability as moderate.

A median reliability of 44.9% is below the target median value using a beta-binomial model, if the intent of the measure is to identify differences between individual facilities. However, the intent of the measure is not to identify differences in performance between individual facilities, but, rather to identify differences from the mean (or threshold) performance value. Thus, we believe the testing reporting in **Section 2b5**, to determine statistically significant and meaningful differences in performance, is a more appropriate test for this measure.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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).*  
**2014 Submission:**

Face validity of the measure score was systematically assessed as follows: after the measure was fully specified, patient-level data and measure specifications were sent to each of the facilities for whom a score was calculated (via individual reports generated during a “dry-run” period). Facilities were provided with an opportunity to review both the specifications and calculations, and to report any concerns regarding the specifications. The validity assessment included the 3,680 facilities in the HOQR Program.

**2016 Submission:**

CORE and Lewin systematically assessed face validity of the measure through survey of the TEP. Composition of the TEP is described in **Section Ad.1**. Eleven TEP members responded to the survey. Respondents included insurers/purchasers, clinicians, management or administration, patients/patient advocates, and caregivers. Prior to responding to questions related to measure-score and data-element face validity, CORE and Lewin provided TEP members detailed measure specifications to support evaluation of the measure’s face validity.

CORE and Lewin posed the following questions and statements related to measure-score face validity:

1. Does NQF #0514 capture the most appropriate and prevalent types of antecedent conservative therapy available through claims data?
2. The measure helps assess the inappropriate use of MRI lumbar-spine tests. Do you agree?

CORE and Lewin collected responses question one using a scale or *yes*, *not sure or do not know*, and *no* and responses to question two using a five-point Likert scale: *strongly agree*, *agree*, *undecided/do not know*, *disagree*, and *strongly disagree*.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
Results of the face-validity survey indicate that a diverse group of stakeholders, a majority of whom were not involved in the measure’s development, support the validity of the measure, including the identification of antecedent conservative therapy.

**2014 Submission:**The results of the assessment of face validity through the dry-run reporting indicate that an independent group of experts (i.e., those different from those who advised on measure development) did not have concerns with the specifications for the measure. Additionally, in an ongoing opportunity for comment on the specifications, clinicians and administrators have not expressed concern regarding the implementation of the specifications.

**2016 Submission:**

*Does NQF #0514 capture the most appropriate and prevalent types of antecedent conservative therapy available through claims data?*

|  |  |  |
| --- | --- | --- |
| **Response Option** | **Response (%)** | **Response (#)** |
| Yes | 72.7 | 8 |
| Not Sure or Do not Know | 9.1 | 1 |
| No | 18.2 | 2 |

Similarly, TEP members indicated that the measure helped to assess the rate of inappropriate use of MRI lumbar spine studies:

*The measure helps assess the inappropriate use of MRI lumbar-spine tests. Do you agree?*

|  |  |  |
| --- | --- | --- |
| **Response Option** | **Response (%)** | **Response (#)** |
| Strongly Agree | 36.4 | 4 |
| Agree | 45.5 | 5 |
| Undecided | 9.1 | 1 |
| Disagree | 0.0 | 0 |
| Strongly Disagree | 0.0 | 0 |
| Do Not Know or Not Applicable | 9.1 | 1 |

**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?*)

**2014 Submission:**

Based on the acceptance of the specifications by facilities in the HOQR Program, the measure is assumed to have face validity. Additionally, the data were considered to have face validity as representing services rendered by the hospital. Additional data testing was not involved.

**2016 Submission:**

TEP evaluation demonstrates face validity of NQF #0514 at the measure-score level. In the primary assessment of face validity of the measure performance score (face-validity question #2), 81.8% of respondents agreed or strongly agreed with the face validity of the measure calculation and believe that the measure helps assess the inappropriate use of MRI lumbar-spine tests. As indicated above, TEP members also responded that the measure captures the most appropriate and prevalent types of antecedent conservative therapy available through claims data.

The results of the TEP survey indicate that the measure has strong face validity. Based on NQF guidance, the strong face validity demonstrated through survey of the TEP merits a rating of “moderate” for the validity criterion. Additionally, claims data have the advantage of being largely error free since CMS audits data fields used in determining payment, further supporting the face validity of the measure. Finally, stakeholders have raised no concerns regarding the face validity of the measure over the six years of public reporting.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**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*).  
**2014 Submission:**

The measure exclusions were tested to determine the prevalence of each exclusion, by facility, and at an aggregate level. Each category of exclusions was also tested to determine the effect on facility performance scores. The analysis tested the following categories of measure exclusions using data from the Medicare FFS 2011 SAF file:

* Lumbar spine surgery (90 day look back)
* Trauma (45 day look back)
* Cancer (12 month look back)
* IV drug abuse (12 month look back)
* Neurologic impairment (12 month look back)
* HIV (12 month look back)
* Unspecified immune deficiencies (12 month look back)
* Intraspinal abscess (No look back)

**2016 Submission:**

CORE and Lewin tested measure exclusions to determine the prevalence of each exclusion, by measured entity, and at an aggregate level. We also tested the effect of all exclusions to determine the total effect of measure exclusions on performance, both by reporting summary statistics and by calculating a spearman rank correlation coefficient. The analysis tested the following categories of measure exclusions in 2013 performance data:

* Cancer
* Congenital spine and spinal cord malformations
* Inflammatory and autoimmune diseases
* Infectious conditions
* Spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage
* Spinal cord infarction
* Neoplastic abnormalities
* Treatment fields for radiation therapy
* Spinal abnormalities associated with scoliosis
* Syringohydromyelia
* Postoperative fluid collections and soft tissue changes
* Trauma
* IV drug abuse
* Neurologic impairment
* HIV
* Unspecified immune deficiencies
* Intraspinal abscess
* Surgery within last 90 days

Currently, the measure excludes patients with any one of the above-listed conditions.

**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*)  
**2014 Submission:**

The measure contractor examined overall frequencies and proportions of the studies excluded for each exclusion criterion in all MRI lumbar spine studies for a sample of 4,202 facilities from the 2011 Medicare FFS SAF file. The initial cohort included 536,583 MRI lumbar spine studies. The final cohort included 312,671 MRI lumbar spines studies. The total number of exclusion occurrences exceeded the number of studies excluded because a single patient might meet multiple exclusion criteria.

|  |  |  |  |
| --- | --- | --- | --- |
| **Exclusion** | **Overall Occurrence (N)** | **Overall Occurrence (%)** | **Distribution Across Hospitals, 25th, 50th, 75th Percentile (%)** |
| Lumbar Spine Surgery | 6,136 | 1.14% | 0, 0, 1.4 |
| Trauma | 59,360 | 11.06% | 6.8, 10.2, 14.2 |
| Cancer | 142,104 | 26.48% | 17.9, 23.5, 29.5 |
| IV Drug Abuse | 11,185 | 2.08% | 0, 1.3, 3.0 |
| Neurologic Impairment | 43,689 | 8.14% | 4.4, 7.3, 10.8 |
| HIV | 2,278 | 0.42% | 0, 0, 0 |
| Unspecified Immune Deficiencies | 890 | 0.17% | 0, 0, 0 |
| Instraspinal Abscess | 197 | 0.04% | 0, 0, 0 |

Additionally, descriptive statistics were calculated for the measure scores of each facility, with and without the exclusion codes, as part of the specifications. Facility scores are noticeably higher without the exclusions, although the standard deviation is smaller.

|  |  |  |
| --- | --- | --- |
| **Descriptive Statistic** | **With Exclusions (%)** | **Without Exclusions (%)** |
| Min | 0 | 0 |
| 10% | 23.0 | 50.9 |
| 25% | 30.3 | 57.8 |
| 50% | 36.4 | 63.1 |
| Mean | 37.4 | 63.5 |
| 75% | 43.8 | 68.8 |
| 90% | 54.5 | 76.9 |
| Max | 100 | 100 |
| Std. Dev. | 15.6 | 12.7 |

The descriptive statistics reported here differ from the publicly-reported statistics, as the mean calculation is not weighted, and the calculation includes some facilities for whom measure value are not publicly reported.

**2016 Submission:**

CORE and Lewin examined overall frequencies and proportions of denominator cases excluded for each exclusion, among all MRI lumbar-spine studies, for a sample of 2,569 facilities meeting the minimum case count requirements in 2013, imposing no measure exclusions. The initial patient population included 521,460 MRI lumbar-spine studies. The total number of exclusion occurrences exceeded the

number of excluded cases because a single beneficiary might meet multiple exclusion criteria.

| **Facilities** | | | | | |
| --- | --- | --- | --- | --- | --- |
| *Exclusion* | *Overall Occurrence (N)* | *Overall Occurrence (%)* | *Distribution Across Facilities (%)* | | |
| *25th* | *50th* | *75th* |
| Cancer | 139,768 | 26.80 | 20.20 | 24.56 | 29.66 |
| Congenital spine and spinal cord malformations | 76,430 | 14.66 | 8.70 | 12.43 | 17.50 |
| Inflammatory and autoimmune diseases | 64,483 | 12.37 | 8.96 | 11.52 | 14.29 |
| Infectious conditions | 3,484 | 0.67 | 0.00 | 0.39 | 1.02 |
| Spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage | 15,028 | 2.88 | 1.24 | 2.22 | 3.59 |
| Spinal cord infarction | 1,175 | 0.23 | 0.00 | 0.00 | 0.19 |
| Neoplastic abnormalities | 1,957 | 0.38 | 0.00 | 0.00 | 0.49 |
| Treatment fields for radiation therapy | 856 | 0.16 | 0.00 | 0.00 | 0.00 |
| Spinal abnormalities associated with scoliosis | 91,325 | 17.51 | 10.59 | 15.31 | 20.93 |
| Syringohydromyelia | 1,086 | 0.21 | 0.00 | 0.00 | 0.21 |
| Postoperative fluid collections and soft tissue changes | 4,157 | 0.80 | 0.00 | 0.58 | 1.20 |
| Trauma | 59,072 | 11.33 | 8.27 | 10.79 | 13.89 |
| IV drug abuse | 15,193 | 2.91 | 1.22 | 2.31 | 3.93 |
| Neurologic impairment | 40,797 | 7.82 | 5.14 | 7.14 | 9.71 |
| HIV | 2,122 | 0.41 | 0.00 | 0.00 | 0.47 |
| Unspecified immune deficiencies | 1,063 | 0.20 | 0.00 | 0.00 | 0.18 |
| Intraspinal abscess | 428 | 0.08 | 0.00 | 0.00 | 0.00 |
| Surgery within last 90 days | 6,114 | 1.17 | 0.00 | 0.77 | 1.56 |
| **All Current Exclusions** | 328,901 | 63.07 | 55.13 | 61.42 | 67.74 |

Additionally, we calculated descriptive statistics for the measure scores of each facility, with and without exclusions.

| **Facilities** | | |
| --- | --- | --- |
| *Descriptive Statistic* | *With Exclusions (%)* | *No Exclusions (%)* |
| Minimum | 15.52 | 11.11 |
| Maximum | 76.00 | 69.23 |
| Mean | 37.35 | 32.54 |
| Standard Deviation | 7.34 | 6.91 |
| 25th Percentile | 32.53 | 28.02 |
| 50th Percentile | 36.84 | 31.69 |
| 75th Percentile | 41.27 | 35.85 |

**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*)  
**2014 Submission:**

The overall frequency of each exclusion varies. The frequency of trauma, cancer, and neurologic impairment is high (each >8%), and there is a substantial range of frequencies for these exclusions across facilities. Exclusion of these procedures is necessary, as inclusion would noticeably bias facility scores.

The distribution across hospitals is narrower and prevalence is lower for the other exclusions (including lumbar spine surgery, drug abuse, HIV, intraspinal abscess, and unspecified immune deficiency), suggesting that the occurrence is more random, and likely would not bias performance results. The measure contractor believes, however, that these exclusions should be retained for the following reasons:

* Lumbar Spine Surgery: The 2011 ACR appropriateness criteria on low back pain finds the use of MRI lumbar spine to be appropriate for candidates for surgery to address low back pain or radiculopathy, and for patients with prior lumbar surgery (Davis, 2011). The measure contractor has included lumbar spine surgery as an exclusion to align the measure specifications with the ACR guideline.
* Drug Abuse: This measure exclusion is supported by the Institute for Clinical Systems Improvement 2012 guideline on adult acute and sub-acute low back pain, the Michigan Quality Improvement Consortium 2012 guideline on the management of acute low back pain, and the 2010 University of Michigan Health Systems guideline on acute low back pain, all of which list drug abuse as a red flag condition for which a MRI lumbar spine study may be appropriate (Goertz, 2012; Michigan Quality Improvement Consortium, 2012; University of Michigan Health System, 2010).
* HIV: The 2012 Michigan Quality Improvement Consortium guideline on acute low back pain and the 2010 University of Michigan Health Systems guideline on acute low back pain list HIV as a red flag condition for which a MRI lumbar spine study may be appropriate (Michigan Quality Improvement Consortium, 2012; University of Michigan Health System, 2010). Four additional guidelines list infection or immunosuppression as a red flag condition (American Academy of Neurology, 2013; Davis, 2012; Goertz, 2012; Work Loss Data Institute, 2011). The measure contractor has maintained HIV as an exclusion to align the measure specifications with these clinical guidelines.
* Intraspinal Abscess: Inclusion of intraspinal abscess as a measure exclusion is in accordance with the 2012 ACR Practice Guideline for the performance of MRI of the adult spine (American Academy of Radiology, 2012).
* Unspecified Immune Deficiency: Immune suppression is listed as a red flag condition for which MRI of the lumbar spine may be appropriate, according to the following guidelines: 2012 Michigan Quality Improvement Consortium Guideline on the management of acute low back pain, 2010 University of Michigan Health Systems guideline on acute low back pain, 2011 ACR Appropriateness Criteria for low back pain (Goertz, 2012; Michigan Quality Improvement Consortium, 2012; University of Michigan Health System, 2010). The measure contractor has retained unspecified immune deficiency as an exclusion category, in accordance with these guidelines.

The necessity of the exclusion codes is further indicated by the comparison of the descriptive statistics for the facility performance, with and without the exclusion codes. The introduction of the exclusion codes reduced the mean facility score by 26.1%.

*References*

American Academy of Neurology. Practice parameters: Magnetic resonance imaging in the evaluation of low back syndrome (Summary statement). Neurology. 2013; 44:767-770.

American Academy of Radiology. Practice Guideline for the Performance of MRI of the Adult Spine. Reston (VA): American College of Radiology (ACR). 2012.

Davis PC, Wippold FJ II, Cornelius RS. Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria® low back pain. [online publication]. Reston (VA): American College of Radiology (ACR). 2011.

Goertz M, Thorson D, Bonsell J, et al. Institute for Clinical Systems Improvement (ICSI). Adult acute and subacute low back pain. Bloomington (MN): ICSI. 2012

Michigan Quality Improvement Consortium. Management of acute low back pain. Southfield (MI): Michigan Quality Improvement Consortium. 2012.

University of Michigan Health System. Acute low back pain. Ann Arbor (MI): University of Michigan Health System. 2010.

Work Loss Data Institute. Low back - lumbar & thoracic (acute & chronic). Corpus Christi (TX): Work Loss Data Institute. 2011.

**2016 Submission:**

The frequency of excluded cases varied substantially across facilities (IQR: 12.61%). Median performance also changes substantially by applying the exclusion conditions. Median performance increases by 5.15% for facilities after applying measure exclusions. Based on the variance in frequency of measure exclusions, as well as the effect on performance scores, measure exclusions are necessary to prevent unfair distortion of facility results.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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***](#section2b5)***.***

**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**.   
**2014 Submission:**

Risk adjustment was determined not to be necessary during measure specification development, as guidelines did not indicate further need for case mix adjustments. During the five years of public reporting for this measure, neither risk-adjustment nor case mix adjustment have been considered or requested.

**2016 Submission:**

This measure is a process measure for which risk adjustment or risk stratification are not necessary. We determined risk adjustment and risk stratification were not appropriate based on the measure evidence base and the measure construct. During the measure development and maintenance process, we performed an annual review of the literature, which included a scan for potential patient subpopulations for which there are differences in the clinical decision to perform MRI lumbar-spine studies for patients with low back pain absent red flag conditions; this review identified no clear evidence of an empirical relationship between SDS and facility-level measure performance.

In addition to the evidence gathered from the literature, stakeholder feedback obtained during implementation and public reporting has not identified concerns related to SDS factors or the need for risk adjustment. This supports the conceptual model upon which the measure is based. As a process-of-care measure, SDS factors should not influence the decision to image a patient with low back pain; rather, adjustment would risk masking such important inequities in care delivery. Variation across patient populations is reflective of differences in the quality of care provided to the disparate patient population included in the measure’s denominator.

**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*).

**2014 Submission:**

Risk adjustment was determined not to be necessary, as guidelines did not indicate further need for case mix adjustment.

**2016 Submission:**

This measure is a process measure for which risk adjustment or risk stratification are not necessary. We determined risk adjustment and risk stratification were not appropriate based on the measure evidence base and the measure construct.

**2b4.4a. What were the statistical results of the analyses used to select risk factors?  
2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

This measure is a process measure for which risk adjustment or risk stratification are not necessary. We determined risk adjustment and risk stratification were not appropriate based on the measure evidence base and the measure construct.

**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)**

**2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

This measure is a process measure for which risk adjustment or risk stratification are not necessary. We determined risk adjustment and risk stratification were not appropriate based on the measure evidence base and the measure construct.

**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***](#question2b49)

**2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

CORE and Lewin did not perform risk adjustment or stratification.

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:   
2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

CORE and Lewin did not perform risk adjustment or stratification.

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
**2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

CORE and Lewin did not perform risk adjustment or stratification.

**2b4.8. Statistical Risk Model Calibration-Risk decile plots or calibration curves**:  
**2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

CORE and Lewin did not perform risk adjustment or stratification.

**2b4.9. Results of Risk Stratification Analysis**:

**2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

CORE and Lewin did not perform risk adjustment or stratification.

**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*)  
**2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

CORE and Lewin did not perform risk adjustment or stratification.

**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*)

**2014 Submission:**

This measure does not use risk adjustment or stratification.

**2016 Submission:**

CORE and Lewin did not perform risk adjustment or stratification.

**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).*   
**2014 Submission:**

One impediment to achieving high levels of precision and accuracy at the facility level is small case counts. This is an issue for many facilities identified in the data, as they do not perform a high volume of services contained within the measure’s specifications. In the situation where a facility provides only a handful of the relevant services that are eligible for this measure, the results of the measure may be significantly impacted and skewed by one or two cases. Minimum case count requirements were developed for each facility in order to assure a 90% confidence level for the observed facility rate.

There are two different processes for determining required case counts depending on whether the facility rate is less than 0.05 or greater than 0.95 (i.e., towards the end of the range of possible rate values), or somewhere between 0.05 and 0.95 (inclusive). Each process has three steps: (1) determine reasonable levels of precision; (2) determine the level of confidence to be required for the measures; and, (3) calculate the case counts needed to meet the precision requirements. For facility rates less than 0.05 or greater than 0.95, the case count needed to attain the required precision was calculated to be 45 cases. For facility rates between 0.05 and 0.95, the case count needed to attain the required precision ranges from 31 to 67 cases. For more details on the minimum case count requirements determinations, please see the supplemental materials.

Prior to the application of the minimum case count and additional public reporting requirements, the measure contractor also tested the statistical significance of the difference between facility performance scores and the mean performance value. For the 2010 data, 18,429 facilities had more than one denominator procedure in the FFS outpatient SAF file.[[1]](#footnote-1) For each facility, the facility performance score and standard deviation was calculated. The same process was performed for the 2011 data, which included 18,560 facilities.

Methodology explaining the minimum case count calculations for this measure can be found at <https://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228889854907&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3D2012_OIE_MCC.pdf&blobcol=urldata&blobtable=MungoBlobs>.

**2016 Submission:**

Among measured entities that perform only a handful of MRIs for patients with low back pain, one or two cases could significantly influence and/or skew the results of the measure. Therefore, CMS applies minimum case count requirements before reporting performance scores for facilities. The minimum case count requirements applied for this measure and other imaging efficiency measures assure a 90% confidence level for the observed rate. We applied this approach to OP-8 in the facility settings.

For OP-8, we use two different processes for determining required case counts depending on whether the performance rate is less than 0.05 or greater than 0.95 (i.e., towards the end of the range of possible rate values), or somewhere between 0.05 and 0.95 (inclusive). Each process has three steps: (1) determine reasonable levels of precision; (2) determine the level of confidence to be required for the measures; and, (3) calculate the case counts needed to meet the precision requirements. For performance rates less than 0.05 or greater than 0.95, we calculated the case count needed to attain the required precision to be 45 cases. For performance rates between 0.05 and 0.95, the case count needed to attain the required precision ranges from 31 to 67 cases. This composite process for setting the minimum case count requirements optimizes precision while also maximizing the number of reporting hospitals.

A more detailed presentation of the methodology explaining the minimum case count calculations for this measure is included in the NQF #0514 (OP-8) measure report posted at <http://qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228695266120> on page 24.

Following the application of the minimum case count, CORE and Lewin examined differences in performance, calculating results (performance scores) for 1,616 facilities. CORE and Lewin computed a 95% confidence interval for each provider’s score; if it did not contain the mean facility, the facility is identified as better than or worse than average.

**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*)  
**2014 Submission:**

Of the 18,429 facilities in 2010 with more than one denominator procedure, 9,385 (50.93%) had a performance value that was statistically significantly different from the weighted mean (or benchmark value). Statistically meaningful difference was defined as when the measure mean (or benchmark value) fell outside of the confidence interval (± 1.96 standard deviations) for the facility score. In 2011, 4,047 (21.80%) of facilities had a performance value that was statistically significantly different from the weighted mean (or benchmark value).

**2016 Submission:**

Below is a distribution of 2013 performance scores for facilities.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Std. Dev.** | **Min.** | **10th Percent** | **Lower Quartile** | **Median** | **Upper Quartile** | **90th Percent** | **Max.** |
| 38.57 | 7.39 | 18.18 | 30.06 | 33.33 | 37.93 | 42.66 | 48.35 | 72.60 |

**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?*)  
**2014 Submission:**

Analysis of the 2010 data, and the subsequent rate of identification of statistically different performance for 50.93% of measured entities, demonstrates the ability of the measure specifications to identify a statistically significant difference in performance from the mean value. The measure contractor purports that the reduced rate of identification of statistically different performance than the mean in 2011 (in contrast to the weighted mean[[2]](#footnote-2) value, which has remained relatively constant), is indicative of the effectiveness of the measure as a benchmarking tool. The reduced rate of identification may be reflective of a convergence on the mean (or benchmark) value.

**2016 Submission:**

The measure is able to detect statistically better and worse performance between facilities. The facility performance scores ranged from 18.18% to 72.60%, with a median of 37.93%. Fifty percent of facilities fell within the interquartile range of 33.33% to 42.66%. The mean ± SD facility performance score was 38.57% ± 7.39%.This analysis indicated that the measure is able to identify statistically significant and clinically meaningful differences in performance for facilities.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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*)

**2014 Submission:**

This measure only uses one set of specifications.

**2016 Submission:**

This measure only uses one set of specifications.

**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*)  
**2014 Submission:**

This measure only uses one set of specifications.

**2016 Submission:**

This measure only uses one set of specifications.

**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*)  
**2014 Submission:**

This measure only uses one set of specifications.

**2016 Submission:**

This measure only uses one set of specifications.

**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*).  
**2014 Submission:**

The measure contractor does not make any adjustments for missing data. The measure relies on Medicare claims data, which are used for payment purposes for services rendered by a provider. The data undergo prepayment claims analysis and post payment audits, as part of the CMS administrative process. The analytic files used by the measure developer are post-adjudicated claims.

**2016 Submission:**

This measure is calculated from claims data submitted by measured entities for purposes of payment. The administrative claims data used to calculate the measure are maintained by CMS’s Office of Information Services; these data undergo additional quality assurance checks during measure development and maintenance. Thus, the analytic files used for measure testing and measure calculation include post-adjudicated claims, and do not include known missing data.

**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*)  
**2014 Submission:**

The analytic files used by the measure developer are post-adjudicated claims, and do not include missing data.

**2016 Submission:**

As described in **Section 2b7.1**, the analytic files used for measure testing and measure calculation include post-adjudicated claims, and do not include known missing data.

**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*).

**2014 Submission:**

The analytic files used by the measure developer are post-adjudicated claims, and do not include missing data. As such, missing data does not bias the performance results.

**2016 Submission:**

As described in **Section 2b7.1**, the analytic files used for measure testing and measure calculation include post-adjudicated claims, and do not include known missing data. As such, missing data does not bias the performance results.

1. Facilities included in this calculation may not be publicly reported. Following initial facility score calculation, another contractor removes facilities that do not meet public reporting requirements, such as those that are not subject to public reporting due to facility status. [↑](#footnote-ref-1)
2. For questions 2b5.2 and 2b5.3 weighted mean refers to the weighted mean of the publicly reported facility scores in the HOQR program. The weighted mean for 2010 was 31.6% and the weighted mean for 2011 was 31.4%. [↑](#footnote-ref-2)