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

**Measure Number** (*if previously endorsed*)**:** 0052

**Measure Title**: Use of Imaging Studies for Low Back Pain

**Date of Submission:** 3/3/2014

**Type of Measure:**

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

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

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

N/A

**1.3. What are the dates of the data used in testing**? Data element validity testing was performed using data from January 1 to December 31, 2002. Testing of face validity was performed in January and May 2004 and re-evaluated in 2012. Measure score reliability testing was performed using data from January 1 to December 31, 2012.

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

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

**1.5. How many and which 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 Element Validity Testing**

We used field test data for data element validity testing. The field test included three health plans that provided patient-level administrative and medical record data to NCQA for this study. Two plans submitted data for only the commercial population, and one plan submitted data for both its commercial and Medicaid population. They represented several geographic regions of the country, and included network models and staff model health plans. The participating plans provided patient information from administrative data systems for the entire eligible population. They also provided medical record information for a random sample of 150 patients in the eligible population. Medical records came from providers identified on the first claim for low back pain during the measurement period.

**Measure Score Reliability Testing**

We used HEDIS data for measure score reliability testing using the beta-binomial method. We calculated the measure score reliability from the most recent HEDIS data, which included 180 Medicaid and 409 commercial health plans. The sample included all Medicaid and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

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

**Data Element Validity Testing**

**Table 1**. **Plan Size for Data Element Validity Testing**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plan | Total Enrollment | Patients w/ LBP Episode Sampled for Validity Testing | Percent of Sample with MR Missing | Final Sample for Validity Testing |
| A | 464,637 | 148 | 17.6 | 122 |
| B | 378,909 | 150 | 0 | 150 |
| C | 111,986 | 150 | 0 | 150 |

Table 1 shows the total enrollment for each plan and the sample size used for the medical record review in 2002. Note that Plan C had both a commercial and Medicaid product line. Only one plan (Plan A) had missing medical records (17.6 percent) in the sample of patients randomly selected for the medical record review. Only those patients whose medical record was found were counted as the final sample for validity testing.

## Table 2. Demographics of Patients with Episodes of LBP for Data Element Validity Testing

|  |  |  |
| --- | --- | --- |
|  | Commercial  Percent  (N=21,777) | Medicaid  Percent  (N=504) |
| Sex |  |  |
| Male | 45.9 | 13.5 |
| Female | 54.1 | 86.5 |
| Age |  |  |
| <=20 | 6.0 | 8.5 |
| 21-30 | 24.2 | 33.7 |
| 31-40 | 32.4 | 38.9 |
| 41-50 | 37.4 | 18.9 |

**Measure Score Reliability Testing**

**Table 3.** **HEDIS Plan Size for Reliability Testing**

|  |  |  |
| --- | --- | --- |
| Product Type | Number of Plans | Median Number of Eligible Patients per Plan |
| Commercial HMO | 210 | 835 |
| Commercial PPO | 199 | 2547 |
| Medicaid HMO | 180 | 698 |

In 2012, HEDIS measures covered 107.3 million commercial health plan members and 21.7 million Medicaid HMO members. Data is summarized at the health plan level. Data is stratified by product line (i.e. commercial and Medicaid). Table 3 provides a description of the sample. It includes the number of health plans included in the HEDIS data collection and the median eligible population for the measure across health plans.

**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**.

**Data Element Validity Testing**

We conducted data element validity testing using field test data in 2002 (described above in Tables 1 and 2).

**Measure Score Reliability Testing**

We conducted reliability testing of the measure score using a beta-binomial calculation. This analysis included the entire HEDIS commercial and Medicaid data samples in 2012 (described above in Table 3).

**Systematic Evaluation of Face Validity**

In addition to data element validity testing, we also completed a systematic assessment of face validity. We tested this measure for face validity with five panels of experts. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel members.

* The Musculoskeletal work group includes 10 experts, including representation by health care providers, foundations, and the Centers for Disease Control and Prevention (CDC). This panel initially assessed face validity in 2003, but the Bone Joint MAP replaced this work group and assessed the measure during the last re-evaluation in 2012.
* The Bone Joint MAP includes 10 experts, including representation by health care providers, health plans, and universities.
* The Technical Measurement Advisory Panel includes 12 members, including representation by health plans methodologists, clinicians and HEDIS auditors.
* The HEDIS Expert Coding Panel includes 10 members, including representation by health plans, hospital associations, and advisory groups.
* NCQA’s Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 18 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

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

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

**Reliability Testing of Performance Measure Score**

In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009). The following is quoted from the tutorial which focused on provider-level assessment: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.” This approach is also relevant to health plans and other accountable entities.

Adams’ approach uses a beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

**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*)

**Results of Reliability Testing of Performance Measure Score**

**Figure 1. Overall Reliability Results**

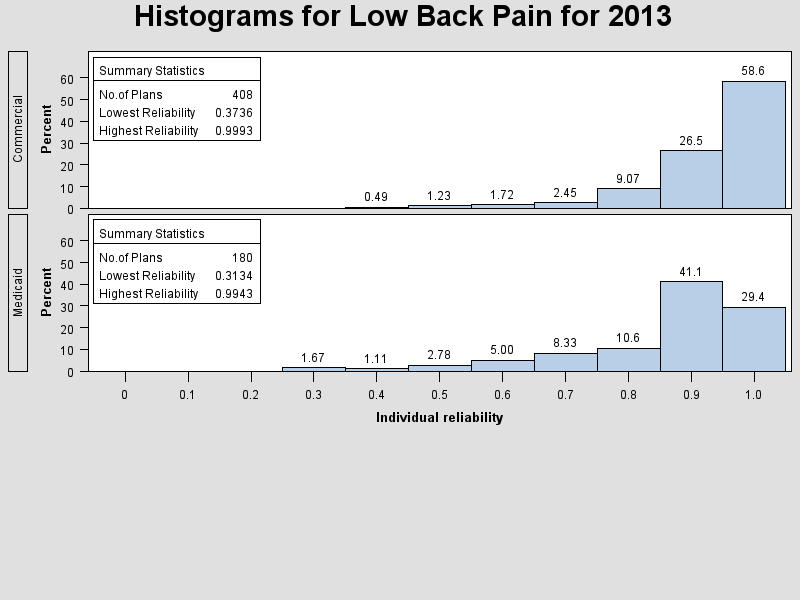
|  |  |  |  |
| --- | --- | --- | --- |
| Commercial | | Medicaid | |
| # of Plans | Reliability Score | # of Plans | Reliability Score |
| 409 | 0.99 | 180 | 0.94 |

**Figure 2. Individual Reliability and Distribution Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Commercial | | Medicaid | |
| Median | 10-90th Percentile | Median | 10-90th Percentile |
| 0.96 | 0.81 - 0.99 | 0.92 | 0.64 - 0.98 |

**Figure 3. Reliability Histograms for Low Back Pain in 2012 across Commercial and Medicaid Plans**

**HISTOGRAMS FOR LOW BACK PAIN FOR 2012**



**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

**Interpretation of Measure Score Reliability Testing**

**Figure 1:** Overall Reliability

The reliability was estimated at 0.99 (commercial) and 0.94 (Medicaid) based on 409 commercial plans and 180 Medicaid plans using the beta-binomial model (Adams, 2009). The beta-binomial method measures the proportion of total variation attributable to a health plan which represents the “signal”. The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan and this is referred to as “noise”. The reliability of the measure is represented as the ratio of signal to noise.

* A score of 0.0 indicates none of the variation (signal) is attributable to the health plan.
* A score of 1.0 indicates all of the variation (signal) is attributable to the health plan.
* A score of 0.7 or higher indicates adequate reliability to distinguish performance between two providers.

**Figures 2 and 3**: Individual Reliability, Distribution, and Histograms

The underlying formulas for the beta-binomial reliability can be adapted to construct a health plan-specific estimate of reliability by substituting variation in the individual health plan’s variation for the average health plan’s variation. As a result, the reliability for some health plans may be more or less than the overall reliability across health plans, just as not everyone who lives in a wealthy neighborhood is wealthy. Figure 2 summarizes the variability of each individual plan’s reliability.

* The median commercial health plan’s reliability at .96 was greater than 0.7, indicating high reliability.
* The median Medicaid health plan’s reliability at .92 was greater than 0.7 indicating high reliability.

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

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

**2b2. VALIDITY TESTING**

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

☒  **Performance measure score**

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

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

**Data Element Validity Testing**

Participating plans provided patient information from administrative data systems and from medical records. Both administrative sources and medical records were used to verify the completeness and accuracy of the administrative data. We assessed validity by comparing the rate of agreement between the administrative codes and the medical records.

**Systematic Assessment of Face Validity**

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

**STEP 1**: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA’s MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

**STEP 2**: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

**STEP 3**: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

**STEP 4:** First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s State of Health Care Quality, Quality Compass or accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA’s experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

**STEP 5**: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publically reported and may be used for scoring in accreditation.

**Step 6:** Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Measures are reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA’s Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year’s data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year’s HEDIS Volume 2.

**ICD-10 Conversion**

The goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

Steps in ICD-9 to ICD-10 Conversion Process

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use GEM to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAPs; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

**Tools Used to Identify/Map to ICD-10**

All tools used for mapping/code identification from CMS ICD-10 website (<http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html>).

GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

**Expert Participation**

The NCQA HEDIS Expert Coding Panel and Bone Joint Measurement Advisory Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panel are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

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

**Results of Data Element Validity Testing**

**Table 4. Medical Record Confirmation of Low Back Pain (LBP) Episode (N=448)**

|  |  |  |  |
| --- | --- | --- | --- |
| Plan | LBP Patients w/ MR\*\* | Medical Record Confirmation Percent | |
| No | Yes |
| A | 122 | 23.0 | 77.0 |
| B | 150 | 5.3 | 94.7 |
| C\* | 150 | 12.0 | 88.0 |

\* Plan C includes both commercial and Medicaid product lines.   
\*\*The total number of patients in the medical record sample with low back pain diagnosis identified in the administrative data.

Table 4 shows the percentage of patients with a claim for low back pain, and a medical record that confirms a low back pain diagnosis. During the field test, a diagnosis of low back pain according to claims data was confirmed by the “gold standard” (i.e. medical record data) in 77–94.7 percent of patients.

## Table 5. Source of Information on Inappropriate Imaging among Low Back Pain (LBP) Patients with Available Medical Records (N=431)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Plan | Percent of Inappropriate Imaging | | | Absence of Inappropriate Imaging  Percent | Total | Percent Agreement  (Admin & MR) + (Neither)\*\* |
| Admin Only | MR Only | Admin & MR | Neither (Admin Nor MR) |
| A | 12.5 | 4.9 | 15.3 | 67.4 | 100.0 | 82.7 |
| B | 18.4 | 0 | 0 | 81.6 | 100.0 | 81.6 |
| C\* | 23.3 | 0 | 0 | 76.7 | 100.0 | 76.7 |

\*Plan C includes both commercial and Medicaid product lines.

\*\*Rate of agreement shows the proportion of patients whose medical record and administrative record agree on the presence or absence of inappropriate imaging.

Table 5 shows the source of information for the inappropriate use of imaging for patients with available medical records and no red-flag diagnoses (exclusions) according to administrative data. During the field test, the identification of imaging (i.e., confirmation of the claims-based presence or absence of inappropriate imaging) was confirmed by the “gold standard” (i.e., medical record data) in 76.7-82.7 percent of patients. For Plan A, inappropriate imaging was identified in claims data but not confirmed by the medical record for 12.5 percent of patients. In addition, the medical record identified an additional 4.9 percent of patients with inappropriate imaging that was not detected by claims. For Plans B and C, inappropriate imaging was identified in claims data but not confirmed by the medical record for 18.4 percent and 23.3 percent of patients, respectively. For these plans, the medical record did not identify any additional inappropriate imaging.

**Results of Face Validity Assessment**

**Step 1:** This measure was developed in 2003 to assess the inappropriate use of imaging for patients with a diagnosis of acute low back pain (LBP) in the absence of indicators of potentially serious spinal pathology or other non-spinal pathology. As a collaborating organization in the American Medical Association (AMA), Joint Commission on Accreditation of Healthcare Organizations (JCAHO), and NCQA Collaboration on Pain Management Performance Measures, NCQA and the Musculoskeletal work group worked together to develop and field test this measure.

**Step 2:** The measure was written in 2003 and field-tested in 2003 using data from 2002. After reviewing field test results, the CPM recommended to send the measure to Public Comment with a majority vote in January 2004.

**Step 3:** The measure was released for Public Comment in 2004 prior to publication in HEDIS. We received and responded to 115 comments on this measure, including 10 organizations that supported the measure. The CPM recommended moving this measure to first year data collection by a majority vote.

**Step 4:** The measure was introduced in 2004. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure to public reporting with a majority vote.

**Step 5**: The measure was re-evaluated in 2012 and reviewed by the Bone Joint Measurement Advisory Panel. The measure was presented to the CPM in September 2012 and the CPM recommended retaining the measure with no changes by majority vote in 2012.

**Conclusion:** The measure was deemed to have the desirable attributes of a HEDIS measure in 2004 and 2012 (relevance, scientific soundness, and feasibility).

**ICD-10 Conversion**

Summary of Stakeholder Comments Received

NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. NCQA received comments from four organizations:

* Support recommendations.
* Questions about select codes.
* Recommended additional codes for consideration.

**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?*)

**Interpretation of Data Element Validity Testing**

Table 4 shows that there is good agreement between claims data and medical record data to identify the diagnosis of low back pain (denominator) with an average confirmation rate of 86.6 percent. Table 5 shows good agreement between claims data and medical record for inappropriate imaging (numerator) with an average agreement rate of 80.3 percent. In addition, Table 5 shows that we found more cases of inappropriate imaging in administrative claims data only than were found in the medical record only. This might be due to the fact that providers may not regularly document the findings of imaging tests in the medical record.

**Interpretation of Systematic Assessment of Face Validity**

These results indicate that the expert panels were in agreement that the measure as specified will accurately differentiate quality across health plans. Our interpretation of these results is that this measure has sufficient face validity.

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

**2b3. EXCLUSIONS ANALYSIS**

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

While the measure does not contain exclusions, the identification of eligible individuals for the denominator includes a negative diagnosis history. Individuals with a history of cancer, recent trauma, intravenous drug abuse, and neurologic impairment are not included in the denominator. We analyzed these “red flags” and others during field testing.

**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*)

**Exclusion Analysis Using Field Test Data**

The aim of testing exclusions in the field test data was to determine how common exclusionary diagnoses (i.e. red flags) would be in the eligible patient population and the impact of these exclusions on denominator sizes and performance rates. Our results (detailed below) show slight differences in performance rates with and without exclusions and across data sources (administrative vs. medical record).

**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*)

**Field Test Data**

## Table 6. Frequency of Exclusionary (Red Flag) Conditions\*

|  |  |  |
| --- | --- | --- |
| Condition | Frequency | Percent of LBP Episodes  (N=22,281) |
| Prior Cancer | 431 | 1.9 |
| Recent Trauma | 198 | 0.9 |
| IV Drug Use | 153 | 0.7 |
| Neurologic Impairment | 156 | 0.7 |
| Recent Infection | 46 | 0.2 |
| Fever | 51 | 0.2 |
| Prolonged Steroid Use | 24 | 0.1 |
| Unexplained Weight Loss | 28 | 0.1 |
| Immunosuppression | 9 | 0.0 |
| Total | 1096 | 4.9 |

\* In administrative data

Table 6 shows the frequency of red-flag conditions (i.e. exclusions) for the imaging of low back pain measure.

## Table 7. Source of Red-Flag/Exclusion Diagnoses (i.e. Justifications for Imaging) Among LBP Patients (N=431)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diagnosis | Rate (Admin Only) | Rate (MR Only) | Rate (Admin & MR) | Rate (Neither Admin nor MR) |
| Recent Trauma | 0.0 | 19.3 | 0.2 | 80.4 |
| Prior Cancer | 0.9 | 0.2 | 0.4 | 98.4 |
| IV Drug Use | 1.1 | 0.0 | 0.0 | 98.9 |
| Neurologic Impairment | 0.9 | 4.0 | 0.0 | 95.1 |
| Recent Infection | 0.2 | 1.8 | 0.0 | 98.0 |
| Fever | 0.2 | 0.4 | 0.0 | 99.3 |
| Unexplained Weight Loss | 0.0 | 0.9 | 0.0 | 99.1 |
| Prolonged Steroid Use | 0.0 | 0.0 | 0.0 | 100.0 |
| Immunosuppression | 0.0 | 0.0 | 0.0 | 100.0 |
| Total | 1.3 | 25.1 | 0.2 | 73.7 |

Table 7 shows the type of exclusion diagnoses (i.e. justifications for imaging) captured for LBP patients with medical records across the different sources of data (administrative, medical records, both, and neither) for the four most common exclusions identified by administrative data.

We have updated Table 7 to reflect all of the exclusions we tested in 2002, since we have added prolonged steroid use, spinal infection, and immunosuppression to the exclusions that are already part of the measure (i.e. recent trauma, prior cancer, IV drug use and neurologic impairment).

**Table 8.** **Measure Rate as Specified for Patients with Available Medical Records (N=431) (lower rate indicates better quality)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Exclusions in Admin Data Only | | Exclusions in Admin or MR Data | |
| Plan | Denominator | Rate | Denominator | Rate |
| A | 144 | 27.8 | 128 | 25.0 |
| B | 141 | 18.4 | 91 | 15.4 |
| C\* | 146 | 23.3 | 102 | 23.5 |

\*Plan C includes both commercial and Medicaid product lines.

Table 8 shows the performance rate (i.e. percentage of inappropriate scans) across plans for patients with exclusions captured through administrative data only and with exclusions captured through either the administrative data or medical record data.

**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*)

~~According to administrative data (Table 6), red-flag conditions (i.e. exclusions) occur in 0.0-1.9 percent of low back pain episodes. Exclusion rates for recent trauma and intravenous drug use increase when plans are able to use both administrative and medical record data (Table 7); however, administrative data is a more valid data source for prior cancer and intravenous drug use. We decided to include the four most common exclusions in the measure, using administrative data only, as this reduces the burden on reporting plans. For plans that are not able to capture recent trauma and intravenous drug use using administrative data, we think the impact on the overall performance rate will be relatively low, demonstrated by Table 8. As part of the field test, we compared measure rates using exclusions identified in administrative data and exclusions identified in either administrative data or medical records (Table 8). The performance rate improved by 2.8-3.0 percentage points for two plans when using both data sources for exclusions; however, Plan C performed worse by 0.2 percent.~~

**Updated analysis including the addition of three exclusions to Tables 6 & 7**

According to the administrative data (Table 6) from 2002, red-flag conditions (i.e. exclusions) occur in 0.0-1.9 percent of low back pain episodes. Using data from both administrative data and medical records (Table 7), we see neurologic impairment, recent infection, recent trauma and unexplained weight loss are more often present in the medical record than administrative data, while IV drug use and prior cancer are more often present in administrative data compared to the medical record. We include eight of these exclusions in our measure, based on the evidence and feedback from stakeholders. The measure is specified using administrative data as the data source, as this reduces the reporting burden on plans. For plans that are not able to capture exclusions using administrative data, we think the impact on the overall performance rate will be relatively low, demonstrated by Table 8. As part of the field test, we compared measure rates using exclusions identified in administrative data and exclusions identified in either administrative data or medical records (Table 8). The performance rate improved by 2.8-3.0 percentage points for two plans when using both data sources for exclusions; however, Plan C performed worse by 0.2 percent.

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

**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)***.***

**N/A**

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

☐ **No risk adjustment or stratification**

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

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

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

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

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors 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*)

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

**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)

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

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

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

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

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

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

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

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

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

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. NCQA calculated an independent sample t-test of the performance difference between randomly selected plans from the top and bottom quartiles of performance.

**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*)

**Table 9. Variation in Performance Across Health Plans (2012)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. # of Patients | Avg. Perf | SD | 10th | 25th | 50th | 75th | 90th | IQR |
| Commercial HMO | 2272 | 75.3 | 6.0 | 66.7 | 70.6 | 75.6 | 79.7 | 82.7 | 9.0 |
| Commercial PPO | 5195 | 74.2 | 5.9 | 67.0 | 69.8 | 74.4 | 78.8 | 81.6 | 9.0 |
| Medicaid HMO | 1119 | 75.6 | 5.7 | 68.3 | 71.5 | 75.2 | 79.3 | 82.3 | 7.8 |

Avg # of patients: the average denominator size across plans

Avg Perf: the average performance rate across plans

IQR: Interquartile range

**Table 10. T-test Between Two Randomly Selected Health Plans (2012)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Plan Rate (25th Percentile) | Plan Rate (75th Percentile) | P-Value |
| Commercial | 68.2 | 81 | <0.001 |
| Medicaid | 70 | 83 | <0.01 |

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

**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?*)

The results above indicate there is a 7-9 percent gap in performance between the 25th and 75th performing plans. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for commercial HMOs which show 9.0 percentage point gap between 25th and 75th percentile plans. For a plan of average eligible population size, this means 207 fewer patients would receive an inappropriate imaging study if treated at a high performing commercial HMO plan compared to a low performing plan.

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

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

***If only one set of specifications, this section can be skipped.* N/A**

**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*)

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

**2b6.3. What is your interpretation of the results in terms of 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*)

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

**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*)

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to 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*)

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to 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*)

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.