TO: Musculoskeletal Standing Committee

FR: NQF Staff

RE: Post-Comment Call to Discuss Public and Member Comments

DA: March 31, 2017

Purpose of the Call

The Musculoskeletal Standing Committee will meet via conference call on Thursday, April 6 from 12:00-2:00 PM ET. The purpose of this call is to:

- Review and discuss comments received during the post-evaluation public and member comment period
- Provide input on responses to the post-evaluation comments
- Re-vote on the Scientific Acceptability criteria for measure #0052
- Determine whether to reconsideration measure #0514 as requested by the measure steward

Standing Committee Actions

1. Review this briefing memo and Draft Report
2. Review and consider the full text of all comments received to the post-evaluation comments (see Comment Table)
3. Be prepared to re-vote on the Reliability and Validity subcriteria for measure #0052 and to consider a re-vote for measure #0514. Complete measure worksheets are provided in Appendix A and the evaluation summaries of the two measures from the draft report are provided in Appendix B.

Conference Call Information

Please use the following information to access the conference call line and webinar:

Speaker dial-in #: (855) 696-3824 (***Committee only***. *No conference code required*.)
Web Link: http://nqf.commpartners.com/se/Rd/Mt.aspx?792275
Public dial-in #: (877) 315-9042 (*No conference code required.*)

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

Background

For this project, the Musculoskeletal Standing Committee evaluated two measures undergoing maintenance review against NQF's standard evaluation criteria during the Committee's late 2016-early 2017 off-cycle activities. The Committee did not recommend the measures for endorsement.

Comments Received

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

Pre-evaluation comments

For this evaluation cycle, the pre-evaluation comment period was open from November 28 – December 9, 2016. No pre-evaluation comments were received.

Post-evaluation comments

The Draft Report was released for public and member comment from February 16 – March 17, 2017. During this commenting period, NQF received ten comments, seven of which were from four member organizations:

Consumers – 0	Professional – 3
Purchasers – 0	QMRI – 0
Health Plans – 1	Providers – 0
Supplier and Industry – 0	Public & Community Health – 0

In order to facilitate discussion, the post-evaluation comments have been categorized by measure. Where possible, NQF staff has proposed draft responses for the Committee to consider. Although all comments are subject to discussion, we will not necessarily discuss each comment and response on the post-comment call. Instead, we will spend the majority of the time considering the major topics and the most significant issues that arose from the comments.

We have included all of the comments that we received in the Comment Table. This table contains the commenter's name, comment, associated measure, and draft responses for the Committee's consideration. Please refer to this comment table to view and consider the individual comments received to each. (Note: Two of the comments received were from one Musculoskeletal Standing Committee member.)

Committee Re-Vote on Measure #0052

#0052: Use of Imaging Studies for Low Back Pain (National Committee for Quality Assurance)

During the evaluation of this measure, the Standing Committee voted against continued endorsement, primarily due to concerns about reliability and validity. For the current submission, the developer revised the specifications but were unable to provide updated testing of the measure. However, they did provide additional information from their 2002 field testing analysis, which provided insight on the ability to identify patients in the recently-added exclusion categories. Committee members noted that the 2002 field testing indicated that a substantial number of patients with trauma or neurologic impairment were not captured using administrative claims data.

Subsequent to the Committee vote, NQF staff re-examined its preliminary rating of the validity subcritieron as "*Insufficient*." Specifically, in the staff preliminary analysis and during the measure evaluation webinar, staff noted that only percentage agreement statistics were provided to show the level of agreement between administrative codes and medical records and that the results provided were not calculated using the newly-specified measure. However, after further consideration of the analysis of the ability to identify the new exclusions in claims only, in medical records only, or in both, staff determined that these data shed light on questions that are addressed in sensitivity/specificity analysis, even though the developer did not provide actual sensitivity/specificity statistics. Thus, staff no longer considers the data element validity testing presented by the developer to be insufficient (staff rating is now "*Moderate*"). Because we have reversed our previous guidance, NQF is asking the Committee to reconsider and re-vote on Reliability and Validity during the post-comment call.

During the evaluation webinar, the developer noted that it might be possible to obtain more recent data for the Committee to consider. Subsequently, the developer provided data from two health plans to

demonstrate the impact of the changes in the measure specifications on the measure denominator, exclusions, and performance rate. (See Appendix C)

NQF received five post-evaluation comments regarding this measure. (Note: One Musculoskeletal Standing Committee member submitted a comment.) Three of the commenters supported the decision of the Committee not to endorse the measure. Commenters emphasized the importance of limiting unnecessary imaging for low back pain, but expressed concerns over the exclusions and the validity of the measure. Two of the commenters supported of the measure.

> **Developer Response**: This measure is not meant to exclude all appropriate reasons for imaging; we have focused on those causes that are evidence-based, are more common and therefore more likely to impact a health plan's performance rate and improve face validity. The exclusions are based on a review of relevant guidelines and evidence, and feedback from several stakeholder and expert groups, and further informed through a 30-day public comment period.

The Committee will re-vote on Reliability and Validity during this call. If the measure passes the Reliability and Validity subcriteria upon re-vote, NQF will ask the Committee to vote on an overall recommendation for or against continued endorsement.

> **Action Item:** After review and discussion of the comments on this measure, the validity analyses initially submitted for the measure, and the additional information provided by the developer, the Committee will re-vote on the Reliability and Validity subcriteria for the measure.

> **Action Item:** If the Committee agrees that the measure passes the Reliability and Validity subcriteria upon re-vote, it will vote on overall suitability for endorsement.

Reconsideration Request

#0514 Lumbar Spine for Low Back Pain (Centers for Medicare and Medicaid Services): Not Recommended

The developer has requested that the Committee reconsider this measure.

During the evaluation webinar, the Committee did not pass the measure on validity. The Committee expressed concerns with the exclusions and the continued inclusion of "elderly" patients in the measure. The Committee also continued to have concerns with using administrative claims data to identify use of antecedent conservative therapies.

NQF received five post-evaluation comments regarding this measure. (Note: One Musculoskeletal Standing Committee member submitted a comment.) Three of the commenters supported the decision of the Committee not to endorse the measure. Two of commenters supported the measure. Commenters emphasized the importance of limiting unnecessary imaging for low back pain, but expressed concerns over the exclusions and the validity of the measure. One commenter—the developer of the measure—formally requested a reconsideration of the measure due to inappropriate application of the evaluation criteria.

> **Developer Comment**:
>
> The Centers for Medicare & Medicaid Services (CMS) has requested a reconsideration of the National Quality Forum (NQF) Musculoskeletal Standing Committee's decision not to recommend NQF #0514, MRI Lumbar Spine for Low Back Pain, for continued endorsement. NQF #0514 was originally endorsed by the Outpatient Imaging Efficiency Steering Committee in October 2008; during the January 6, 2017 review webinar, it did not pass the Validity criterion.

Based on NQF's Measure Evaluation Criteria and Guidance, we believe that NQF #0514 aligns with the moderate validity recommendation from algorithm #3 (Guidance for Evaluating Validity), as it has received in prior evaluations for endorsement. The measure specifications are aligned with the most updated clinical practice guidelines and have strong face validity; additionally, measure testing confirms that threats to validity have been addressed by the exclusion of red-flag conditions. NQF #0514 also passed the Importance and Reliability criteria during endorsement maintenance review. As one Standing Committee member stated during the review webinar, there will always be exceptions in health care, and, as long as the rate of exceptions is low, performance scores will not be impacted and the measure serves its purpose; we believe that, as currently specified, the measure addresses the broader patterns of care.

Please refer to the full text of CMS's reconsideration request for additional detail on why NQF #0514 should maintain its endorsed status (see Appendix D).

Action Item: Based on comments received and the information provided by the developer, would the Committee like to reconsider this measure?

Action Item: If, upon re-vote, the Committee agrees that the measure passes the Validity subcriteria, it will discuss and then vote on Feasibility, Usability and Use, and on overall suitability for endorsement (see Appendix A).

Action Item: If this measure, as well as measure #0052, is recommended for endorsement, discuss potential areas for harmonization for the two measures (see Appendix E).

Appendix A: Measure Worksheets

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0052

Measure Title: Use of Imaging Studies for Low Back Pain

Measure Steward: National Committee for Quality Assurance

Brief Description of Measure: The percentage of patients with a primary diagnosis of low back pain who did not have an imaging study (plain X-ray, MRI, CT scan) within 28 days of diagnosis.

Developer Rationale: This measure assesses the overuse of imaging studies (plain x-ray, MRI, and CT scans) in adults with acute, uncomplicated low back pain. The intent of this measure is to reduce inappropriate imaging for low back pain –
that is, imaging in the absence of "red flags" (indications that back pain is caused by a serious, underlying pathology that would warrant imaging). Inappropriate imaging is problematic because it is not associated with improved outcomes and exposes patients to unnecessary harms such as radiation exposure and further unnecessary treatment (Chou, Fu, Carrino and Deyo, 2009).

Chou R, Fu R, Carrino JA, Deyo RA. 2009. "Imaging strategies for low-back pain: systematic review and meta-analysis." Lancet 373:463-72.

Numerator Statement: Patients who received an imaging study (x-ray, CT, MRI) within the 28 days following a diagnosis of low back pain.

Denominator Statement: All patients 18 years as of January 1 of the measurement year to 50 years as of December 31 of the measurement year with a claim/encounter for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, with a principal diagnosis of low back pain during the Intake Period (January 1 – December 3 of the measurement year).

Denominator Exclusions: Because the intent of the measure is to assess imaging for patients with a new episode of low back pain, exclude patients with a recent diagnosis of low back pain.

Also, exclude any patient who had a diagnosis for which imaging is clinically appropriate. Any of the following meet criteria:

(1) Cancer
(2) Trauma
(3) Recent IV drug abuse
(4) Neurologic impairment
(5) HIV
(6) Spinal infection

(7) Major organ transplant
(8) Prolonged use of corticosteroids

Measure Type: Process

Data Source: Claims (Only)

Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Aug 10, 2009

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Summary of prior review in 2014:

In the prior review, the Committee discussed the Clinical Practice Guideline for the treatment of Adult Acute and Subacute Low Back Pain from the Institute for Clinical Systems Improvement (ICSI), updated November 2012. The ICSI guideline, states "*Clinicians should not recommend imaging (including computed tomography [CT], magnetic resonance imaging [MRI] and x-ray) for patients with non-specific low back pain.*" The Committee questioned the value of the ICSI guideline, noting that only six small randomized controlled trials (RCTs) were used to develop the guideline, and if the limited study populations were representative of all patients especially considering the exclusion of other guidelines and numerous systematic reviews on this topic. The ICSI guideline was graded as strong recommendation with moderate evidence base.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided updated 2015 American College of Radiology (ACR) Appropriateness Criteria: Low Back Pain, based on a systematic review of 12 studies (3 well designed studies, 2 good quality studies, and 7 quality studies that may have design limitations); 18 supporting references also were included in the review. Six clinical variants are described. Procedures for

uncomplicated acute low back pain (LBP) and/or radiculopathy with no "red flags" were rated as 1 or 2 (usually not appropriate), meaning that "the imaging procedure or treatment is unlikely to be indicated in the specified clinical scenarios, or the risk-benefit ratio for patients is likely to be unfavorable". Ratings for procedures used with other variants varied.

Exception to evidence

N/A

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review and grading conducted (Box 3) → QQC present for ICSI guideline (Box 4) → Evidence graded as moderate quality and SRs agree with not recommending imaging for non-specific low back pain→ Moderate

Questions for the Committee:

- *The evidence provided by the developer is updated but directionally the same as that presented for the previous NQF review. Does the Committee agree and so there is no need for repeat discussion and vote on Evidence?*

Preliminary rating for evidence: ☒ Pass ☐ No Pass

The highest possible rating is MODERATE for evidence.

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

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

- Data presented by the developer indicated substantial variation in the rate of appropriate imaging for patients with low back pain across health plans.

  Commercial Rate (HMO and PPO Combined)
  YEAR| MEAN | ST DEV | Min | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
  2014 | 75% | 6% | 54% | 68% | 71% | 75% | 80% | 83% | 9
  2013 | 75% | 6% | 26% | 67% | 70% | 75% | 79% | 83% | 9
  2012 | 75% | 6% | 56% | 67% | 70% | 75% | 79% | 82% | 9
  2011 | 74% | 6% | 45% | 66% | 69% | 74% | 79% | 82% | 9
  2010 | 74% | 6% | 53% | 66% | 70% | 74% | 78% | 81% | 8

  Medicaid Rate (HMO and PPO Combined)
  YEAR| MEAN | ST DEV | Min | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
  2014 | 75% | 6% | 55% | 68% | 71% | 75% | 78% | 83% | 7
  2013 | 76% | 5% | 58% | 68% | 72% | 75% | 78% | 84% | 6
  2012 | 76% | 6% | 58% | 68% | 72% | 75% | 79% | 82% | 8
  2011 | 76% | 5% | 62% | 70% | 72% | 76% | 79% | 82% | 7
  2010 | 75% | 6% | 58% | 67% | 72% | 76% | 80% | 82% | 8

- In addition, the developer presented data showing geographic variation in performance.

Disparities

- The developer cited 1 study from the Department of Veterans Affairs, which found significantly higher rates of MRI in younger adults compared to older adults and significantly lower rates in blacks compared to whites.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of other evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

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

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

Data source(s): Claims

Specifications:

- The level of analysis is the health plan and is specified for use in the clinician office/clinic, emergency department, and ambulatory urgent care settings. A higher score indicates better quality.
- Specifications have been updated since the last review.
- The numerator for this measure includes patients who received an imaging study (x-ray, CT, MRI) within the 28 days following a diagnosis of low back pain.
- The denominator includes patients ages 18-50 with a claim/encounter for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, with a principal diagnosis of low back pain during the Intake Period (January 1 – December 3 of the measurement year).
- Exclusions include patients with:
  - a diagnosis of uncomplicated low back pain during the 6 months prior to the Index Episode Start Date
  - Cancer
  - Trauma
  - Recent IV drug abuse
  - Neurologic impairment
  - HIV
  - Spinal infection
  - Major organ transplant
  - Prolonged use of corticosteroids
- The CPT, ICD-9, ICD-10, and other codes used to identify patients in the numerator and denominator are included in the value sets excel attachment.
- The calculation algorithm is stated in S.18 and appears straightforward.
- This measure is not risk-adjusted.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*
- *Are the exclusions listed appropriate for this measure?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:
For the previous evaluation, NCQA provided reliability statistics including a signal-to-noise analysis using HEDIS health plan performance data from 2012. The Committee was satisfied with the developer's interpretation of the measure score reliability testing.

Describe any updates to testing
No updated testing was provided.

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing
Reliability was assessed using data obtained from 180 Medicaid and 409 commercial health plans. A beta-binomial method was used to determine the ratio of signal to noise. A signal-to-noise analysis quantifies the amount of variation in a performance measure that is due to true differences (i.e., signal) as opposed to random measurement error (i.e., noise). Results will vary based on the amount of variation between the providers and the number of patients treated by each provider. This method results in a reliability statistic that ranges from 0 to 1. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences. A value of 0.7 often is regarded as a minimum acceptable reliability value. This is considered an appropriate test for measure score reliability.

Results of reliability testing

Overall Reliability Results

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

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

Guidance from the Reliability Algorithm
Precise specifications (Box 1) → no empirical reliability testing for updated version of the measure (box 2) → Insufficient

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

Rationale: The changes to the population being measured in the current version of the specifications (i.e**.,** including physical therapy and telehealth visits and the inclusion of additional exclusions) meets NQF's definition of a material change. NQF requires that testing be conducted for the measure as specified; therefore, updated testing using the revised specifications is needed. The measure would be eligible for a high or moderate rating if updated testing is provided.

If additional data element validity testing/analysis is conducted, the requirements for data element reliability testing would be met and the highest possible rating would be MODERATE for reliability.

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

2b1. Validity: Specifications

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

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

From the previous evaluation: The Committee raised concerns regarding the lack of "red flag" exclusions for conditions that potentially indicate a serious health condition (e.g., unexplained weight loss, insidious onset; unexplained fever; history of urinary or other infection; immunosuppression; diabetes mellitus; prolonged use of corticosteroids; osteoporosis; prior lumbar spine surgery). Some Committee members found the lack of exclusion of these conditions a significant threat to validity and ultimately, agreed the measure did not meet the validity criterion. The CSAC noted the developer's assertion that the frequency of the exclusions suggested by the Committee was very low, and that not including them would not distort the measure. The CSAC requested that NCQA be given time to address the Committee's concerns and the measure brought back for reconsideration.

Question for the Committee:

- *The developer has revised the measure and now excludes individuals with several of the "red flag" conditions. Do you agree that the changes to the specifications are consistent with the evidence to the extent possible?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

- For the prior evaluation, the developer provided an assessment of face validity and data element validity testing.

Describe any updates to validity testing: No updated testing was provided.

SUMMARY OF TESTING

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

Method of validity testing of the measure score:

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

Validity testing method:

- 2004 Face Validity (note: face validity was re-evaluated in 2012): The measure was assessed for face validity with input from 5 NCQA expert panels. NCQA uses a process called the HEDIS measure life cycle to assess face validity. This process includes field testing the measure, soliciting public comment, a one year data collection period and approval from an expert panel before using the measure for scoring in accreditation and public reporting. Note: According to NQF requirements, 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. It is unclear whether or how the expert panel, commenting, and dry-run processes fulfill this requirement and thus the developer should provide additional explanation.
- 2002 Critical Data Element Validity Testing: Validity at the data element-level was tested using data from 30 health plans and 150 patients by comparing the presence of administrative claims codes for a new diagnosis low back pain (required to calculate the denominator) and for the identification of imaging (required to calculate the numerator) to documentation in the medical record, which is considered the "gold standard". Testing results for data elements used for exclusions were not provided.

Validity testing results:

- Face Validity: Based on the expert panel discussions, commenting, and dry run, the results, the developer states that *the "measure was deemed to have the desirable attributes of a HEDIS measure in 2004 and 2012 (relevance, scientific soundness, and feasibility)" and that "These results indicate that the expert panels were in agreement that the measure as specified will accurately differentiate quality across health plans."*
- Critical Data Element Validity Testing:

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

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 Only	MR Only	Admin & MR	Neither (Admin Nor MR)		(Admin & MR) + (Neither)**
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

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- In the previous review, the Committee questioned why certain "red flag" conditions were not excluded from the measure. These "red flag" conditions included unexplained weight loss, insidious onset; unexplained fever; history of urinary or other infection; immunosuppression; diabetes mellitus; prolonged use of corticosteroids; osteoporosis; prior lumbar spine surgery. Some Committee members found the lack of exclusion of these conditions a significant threat to validity.
- The developer provided additional analysis from their 2002 field testing analysis, including the addition of three exclusions that are already a part of the measure: prolonged steroid use, spinal infection, and immunosuppression.
- The developer stated that according to the administrative data 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, 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. Eight of these exclusions are included in the measure based on the evidence and feedback from stakeholders.

Questions for the Committee:

- *Did the developer address concerns regarding exceptions that were noted during the prior review?*
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

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

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

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

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

Question for the Committee:

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

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The developer notes that plans collect this measure using all administrative data sources and NCQA's audit process verifies that plans' measure calculations are not biased due to missing data. However, they do not indicate the extent of missing data nor explain how they verify that measure results are not biased.

Guidance from the Validity Algorithm Specifications somewhat consistent with evidence (Box 1) → Potential threats to validity mostly assessed (Box 2) → empirical testing not conducted at the score level (Box 3) → empirical testing conducted at the data element level (Box 10) → method not appropriate because only percent agreement reported and no testing for exclusion data elements (Box 11) → Insufficient

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

Rationale: Testing for data element exclusions not provided and only percent agreement statistics are given. It is likely that the developer has the needed information from their exclusion analysis to calculate estimate agreement between claims and medical record data. Additional agreement statistics (e.g., kappa values) or sensitivity/specificity statistics likely also can be calculated from the data collected in the field test. Finally, the developer may be able to explain how their validation processes (i.e., work with expert panels, commenting, dry runs) fulfill NQF's face validity requirement to explicitly address whether measure results can distinguish good from poor quality of care.

If the developer conducts additional data element validity testing/analysis or explains how their validation process fulfills NQF's face validity requirements, the highest possible rating would be MODERATE for validity.

Committee pre-evaluation comments

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

Criterion 3. Feasibility

Maintenance measures **– no change in emphasis – implementation issues may be more prominent**

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are in defined fields in electronic claims and generated or collected by and used by healthcare personnel during the provision of care. The data are coded by someone other than person obtaining original information.
- NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

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

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details: The developer reports on multiple uses of the measures:

- Annual State of Health Care Quality Report
- California's Value Based Pay for Performance Program
- Consensus Core Quality Measures Set
- CMS Eligible Professional EHR Incentive Program
- Health Plan Accreditation
- Health Plan Ratings/Report Cards
- HEDIS Accountable Care Organization Accreditation
- Physician Quality Report System
- Quality Compass
- Quality Rating System

Improvement results: From 2012-2014, the average rate has remained around 75% for both Commercial, Medicare and Medicaid plans. Rates in the 10th and 90th percentile scores over the same years were approximately 67% and 83% for commercial plans and 68% and 84% for Medicaid

plans, respectively. Data going back to 2005 reveals that the average performance scores have remained relatively unchanged, with averages ranging from 73-79 percent for both plan types from 2005-2014.

Unexpected findings (positive or negative) during implementation: The developer did not report any unexpected findings.

Potential harms: None have been identified.

Vetting of the measure: NQF has recently added a new subcriterion under Usability and Use: 4d: Vetting of the measure by those being measured and others is demonstrated when:

1) those being measured have been given performance results and data, as well as assistance with interpreting the measure results and data
2) those being measured and other users have been given an opportunity to provide feedback on the measure performance and implementation
3) this feedback has been considered when changes are incorporated into the measure

- The developer's submission does not include the items needed to evaluate whether vetting has been done for this measure. However, the developer notes an auditing process that may meet NQF's requirements for vetting. The developer will be invited to provide additional information during the evaluation meeting.

Feedback : No feedback has been provided on via QPS. The measure was reviewed by MAP for the Medicare Shared Savings Program in 2015. MAP did not support this measure for MSSP because it only applies to patients ages 18-50 years. A gap was noted in measures for overuse of imaging for low back pain in the older population.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*
- *How has the measure been vetted in real-world settings by those being measure or others?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

- **Competing measure:**0514: MRI Lumbar Spine for Low Back Pain (CMS)

Harmonization

- Due to differences in the level of analysis and care settings, the Committee will not be asked to select a best-in-class measure.
- Since the last evaluation, the developers have worked to harmonize the measures, resulting in greater congruence in how low back pain is defined, how cancer exclusions are defined, and in exclusion categories. Areas where the measures are not yet

harmonized include the specific imaging modalities captured by the measure and some of the exclusion categories.

Endorsement + Designation

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

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

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

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because score-level validity testing has not been conducted.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

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

Measure Title: Use of Imaging Studies for Low Back Pain

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

Date of Submission: 3/3/2014

Instructions

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

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

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

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

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

☒ Process: Imaging studies for low back pain

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

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

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

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

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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

The rate in this measure relates to the desired outcome in the following way: Patient is diagnosed with low back pain >>> Health care provider conducts evaluation to characterize severity and cause of low back pain >>> Health care provider and patient discuss whether patient has any "red flags" for which imaging is clinically appropriate >>> If patient does not have any "red flags" and is within 28 days of diagnosis, patient does not receive imaging for low back pain >>> Patient and health care

provider discuss alternative treatment options >>> Patient avoids potentially harmful effects from unnecessary imaging (desired outcome).

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

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

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

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

☐ Other – ***complete section*** *1a.8*

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

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

Guideline #1:

Goertz M, Thorson D, Bonsell J, Bonte B, Campbell R, Haake B, Johnson K, Kramer C, Mueller B, Peterson S, Setterlund L, Timming R. Institute for Clinical Systems Improvement. Adult Acute and Subacute Low Back Pain. Updated November 2012.

Guideline #2:

Patel ND, Broderick DF, Burns J, Deshmukh TK, Fries IB, Harvey HB, Holly L, Hunt CH, Jagadeesan BD, Kennedy TA, O'Toole JE, Perlmutter JS, Policeni B, Rosenow JM, Shroeder JW, Whitehead MT, Cornelius RS, Corey AS, Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria® low back pain. Reston (VA): American College of Radiology (ACR); 2015. 12 p.

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

Guideline #1:

Institute for Clinical Systems Improvement Health Care Guideline for Adult Acute and Subacute Low Back Pain.

Clinicians should not recommend imaging (including computed tomography [CT], magnetic resonance imaging [MRI] and x-ray) for patients with non-specific low back pain *(Strong Recommendation, Moderate Quality Evidence)*. (page 12)

Clinicians should not recommend imaging (including computed tomography [CT], magnetic resonance imaging [MRI] and x-ray) for patients in the first six weeks of radicular pain *(Strong Recommendation, Moderate Quality Evidence)*. (page 29)

Given that low back pain is overall a benign condition, the first task of the evaluation is to identify and address potential red flags that would require further investigation. (page 12)

At each visit, evaluate for presence or absence of red flags and document findings. Red flags include the following:

- Risk factors for cancer including age 50 years old or older with a history of cancer, unexplained weight loss and failure to improve after four to six weeks of conservative low back pain therapy. If all three of these risk factors for cancer are absent, studies suggest that cancer can be ruled out with 100% sensitivity.
- Risk factors for possible spinal infection including intravenous drug use, immunosuppression, urinary infection, fever above 38°C (110.4°F) for greater than 48 hours, and history of tuberculosis or active tuberculosis.
- Signs or symptoms of Cauda Equina Syndrome.
    - New onset of urinary incontinence
    - Urinary retention (if no urinary retention, the likelihood of Cauda Equina Syndrome is less than 1 in 10,000)
    - Saddle anesthesia, unilateral or bilateral sciatica, sensory and motor deficits, and abnormal straight leg raising
- Increased risk factors for fragility fracture.
    - Osteoporosis
    - History of steroid use
    - Immunosuppression
    - Serious accident or injury (fall from heights, blunt trauma, motor vehicle accident) – does not include twisting or lifting injury unless other risk factors are present (e.g., history of osteoporosis)
    - Clinical suspicion of ankylosing spondylitis
    - Drug or alcohol abuse (increased incidence of osteomyelitis, trauma, fracture)
- Unrelenting night pain or pain at rest (increased incidence of clinically significant pathology).
- Consideration of other non-spine origins. (pages 14-15)

Guideline: #2

American College of Radiology (ACR) Appropriateness Criteria: Low Back Pain

- Uncomplicated acute LBP and/or radiculopathy are benign, self-limited conditions that do not warrant any imaging studies.

- MRI of the lumbar spine should be considered for those patient presenting with red flags raising suspicion for serious underlying condition, such as cauda equina syndrome (CES), malignancy, or infection.
- In patients with a history of low-velocity trauma, osteoporosis, or chronic steroid use, initial evaluation with radiographs is recommended.
- In the absence of red flags, first-line treatment for chronic LBP remains conservative therapy with both pharmacologic and nonpharmacologic (eg, exercise, remaining active) therapy.
- If there are persistent or progressive symptoms during or following 6 weeks of conservative management and the patient is a surgery or intervention candidate or diagnostic uncertainty remains, MRI of the lumbar spine has become the initial imaging modality of choice in evaluating complicated LBP. (page 10)

Variant 1: Acute, subacute, or chronic uncomplicated low back pain or radiculopathy. No red flags. No prior management. (page 1)

RADIOLOGIC PROCEDURE	RATING
MRI lumbar spine without contrast	2
X-ray lumbar spine	2
X-ray myelography and post myelography CT lumbar spine	2
Tc-99m bone scan with SPECT spine	2
CT lumbar spine without contrast	2
CT lumbar spine with contrast	2
MRI lumbar spine without and with contrast	2
CT lumbar spine without and with Contrast	1

Variant 2: Acute, subacute, or chronic uncomplicated low back pain or radiculopathy. One or more of the following: low velocity trauma, osteoporosis, elderly individual, or chronic steroid use. (page 2)

RADIOLOGIC PROCEDURE	RATING
X-ray lumbar spine	7
CT lumbar spine without contrast	7
MRI lumbar spine without contrast	7
Tc-99m bone scan with SPECT spine	3
CT lumbar spine with contrast	3
CT lumbar spine without and with contrast	1
X-ray myelography and post myelography CT lumbar spine	1
X-ray discography and post-discography CT lumbar spine	1

Variant 3: Acute, subacute, or chronic low back pain or radiculopathy. One or more of the following: suspicion of cancer, infection, or immunosuppression. (page 3)

RADIOLOGIC PROCEDURE	RATING

MRI lumbar spine without and with contrast	8
MRI lumbar spine without contrast	7
CT lumbar spine with contrast	6
CT lumbar spine without contrast	6
X-ray lumbar spine	5
Tc-99m bone scan whole body with SPECT spine	4
FDG-PET/CT whole body	4
CT lumbar spine without and with contrast	3
X-ray myelography and post myelography CT lumbar spine	3

Variant 4: Acute, subacute, or chronic low back pain or radiculopathy. Surgery or intervention candidate with persistent or progressive symptoms during or following 6 weeks of conservative management. (page 4)

RADIOLOGIC PROCEDURE	RATING
MRI lumbar spine without contrast	8
CT lumbar spine with contrast	5
CT lumbar spine without contrast	5
MRI lumbar spine without and with contrast	5
X-ray myelography and post myelography CT lumbar Spine	5
X-ray lumbar spine	4
Tc-99m bone scan with SPECT spine	4
X-ray discography and post-discography CT lumbar spine	3
CT lumbar spine without and with contrast	3

Variant 5: Low back pain or radiculopathy. New or progressing symptoms or clinical findings with history of prior lumbar surgery. (PAGE 5)

RADIOLOGIC PROCEDURE	RATING
MRI lumbar spine without and with contrast	8
CT lumbar spine with contrast	6
CT lumbar spine without contrast	6
MRI lumbar spine without contrast	6
X-ray myelography and post myelography CT lumbar spine	5
X-ray lumbar spine	5
Tc-99m bone scan with SPECT spine	5
X-ray discography and post-discography CT lumbar spine	5
CT lumbar spine without and with contrast	3

Variant 6: Low back pain with suspected cauda equina syndrome or rapidly progressive neurologic deficit. (page 5)

RADIOLOGIC PROCEDURE | RATING

MRI lumbar spine without contrast | 9

MRI lumbar spine without and with contrast | 8

X-ray myelography and post myelography CT lumbar spine | 6

CT lumbar spine with contrast | 5

CT lumbar spine without contrast | 5

X-ray lumbar spine | 3

CT lumbar spine without and with contrast | 3

Tc-99 bone scan with SPECT spine | 2

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Guideline #1:

The Institute for Clinical Systems Improvement to the guideline assigned a "moderate" grade to the quality of evidence and a "strong" grade to the strength of the recommendation. See table under 1a.4.2. for the grade given to the guideline.

Category	Quality Definition	Strong Recommendation
Moderate Quality Evidence	Further research is likely to have an important impact on the work group's confidence in the estimate of effect and may change the estimate.	The work group is confident that the benefits outweigh the risks, but recognizes that the evidence has limitations. Further evidence may impact this recommendation. This is a recommendation that likely applies to most patients.

Guideline #2:

The American College of Radiology (ACR) Appropriateness Criteria Category Names and Definitions

Rating	Category Name	Category Definition	Disagreement
7, 8, or 9	Usually appropriate	The imaging procedure or treatment is indicated in the specified clinical scenarios at a favorable risk-benefit ratio for patients.	The dispersion of individual ratings from the panel median rating is assessed to determine if there is no disagreement.
4, 5, or 6	May be appropriate	The imaging procedure or treatment may be indicated in the specified clinical scenarios as an alternative to imaging procedures or treatments with a more favorable risk-benefit	

		ration, or risk-benefit ration for patients is equivocal.	When the individual ratings are too dispersed from the panel median (disagreement), "May be appropriate" is the designated rating category.
1, 2, or 3	Usually not appropriate	The imaging procedure or treatment is unlikely to be indicated in the specified clinical scenarios, or the risk-benefit ratio for patients is likely to be unfavorable	

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

Other Institute for Clinical Systems Improvement grades:

Guideline #1:

Category	Quality Definitions	Strong Recommendation	Weak Recommendation
High Quality Evidence	Further research is very unlikely to change the work group's confidence in the estimate of effect.	The work group is confident that the desirable effects of adhering to this recommendation outweigh the undesirable effects. This is a strong recommendation for or against. This applies to most patients.	The work group recognizes that the evidence, though of high quality, shows a balance between estimates of harms and benefits. The best action will depend on local circumstances, patient values or preferences.
Moderate Quality Evidence	See table in 1a.4.3	See table in 1a.4.3	The work group recognizes that there is a balance between harms and benefit, based on moderate quality evidence, or that there is uncertainty about the estimates of the harms and benefits of the proposed intervention that may be affected by new evidence. Alternative approaches will likely be better for some patients under some circumstances.
Low Quality Evidence	Further research is very likely to have an	The work group feels that the evidence	The work group recognizes that there is

	important impact on the work group's confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain.	consistently indicates the benefit of this action outweighs the harms. This recommendation might change when higher quality evidence becomes available.	significant uncertainty about the best estimates of benefits and harms.

Guideline #2:

The rating system is provided in Section 1.a.4.3.

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

Guideline #1:

https://www.icsi.org/_asset/7mtqyr/ReviewingEvidenceUsingGRADE.pdf

Guideline#2:

https://acsearch.acr.org/list

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

☒ Yes → ***complete section*** *1a.7*

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

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

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

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

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

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

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

Complete section 1a.7

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

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

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

Complete section 1a.7

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

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

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

The following evidence review supports the guideline from the Institute for Clinical Systems Improvement.

The evidence review assessed the benefits and harms of conducting imaging studies for patients with acute low back pain who do not have any "red flags." This aligns with the measure, which assesses the proportion of patients with a primary diagnosis of low back pain who did not have an imaging study (plain X-ray, MRI, CT scan) within 28 days of the diagnosis of low back pain. Appropriate treatments within this timeframe for most patients include pain medications, advice to stay active, and reassurance from the health care provider.

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

The Institute for Clinical Systems Improvement assigned a "moderate" grade to the quality of evidence.

Category	Quality Definitions
Moderate Quality Evidence	Further research is likely to have an important impact on the work group's confidence in the estimate of effect and may change the estimate.

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

Other Institute for Clinical Systems Improvement quality of evidence grades:

Category	Quality Definitions
High Quality Evidence	Further research is very unlikely to change the work group's confidence in the estimate of effect.
Low Quality Evidence	Further research is very likely to have an important impact on the work group's confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 1976-2011

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

The Institute for Clinical Systems Improvement reviewed three meta-analyses that included studies regarding inappropriate imaging for patients with non-specific low back pain. The meta-analyses systematically reviewed randomized controlled trials (RCTs). Two of the meta-analyses were specific to imaging strategies for low back pain, while the third focused on interventions for improving the appropriate use of imaging for low back pain. This submission concentrates on the body of evidence found in the two meta-analyses specific to imaging strategies for low back pain. Those two meta-analyses identified the same 6 RCTs that directly supported the guideline.

Five of the RCTs met at least four of eight predefined quality criteria, and were classified as higher quality.

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

Overall, for the two meta-analyses specific to imaging for low back pain, there is high quality evidence supporting the non-use of imaging within 28 days of a low back pain diagnosis for patients presenting without “red flags.” Six randomized controlled trials provide evidence for this guideline.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

This measure intends to reduce the use of inappropriate imaging studies (studies within 28 days of diagnosis for patients without red flags). As an overuse measure, the evidence for this measure needs to

demonstrate that the harms of imaging in the first 28 days for patients without red flags outweigh the benefits of imaging in the first 28 days for patients without red flags. We present evidence below that there is little harm to patients by avoiding the use of imaging for patients without indications of underlying conditions, and there is significant radiation exposure (harm) to patients who receive imaging studies.

One of the meta-analyses cited by the Institute for Clinical Systems Improvement guideline concluded that, "immediate, routine lumbar-spine imaging in patients with low-back pain and no features suggesting serious underlying conditions did not improve clinical outcomes compared with usual clinical care without immediate imaging. Results were limited by small numbers of trials for some analyses, but seemed consistent for the primary outcomes of pain and function, and for quality of life, mental health, and overall improvement. Data for patient satisfaction could not be pooled, but showed no clear difference. In addition to non-significance, pooled estimates were small or close to zero and, in some cases, slightly favored the non-imaging strategy. This result suggests that, even if statistical power could be increased by other trials, clinically important benefits from routine lumbar imaging are unlikely, assuming that future results are similar to those currently available" (Chou, 2009).

The figure below represents the overall impact of immediate lumbar imaging versus usual clinical care without immediate imaging. The graph, shown on the right, demonstrates that risk ratios less than 1 indicate that the overall results for pain, function, quality of life and mental health favored non-use of immediate lumbar imaging.

Figure 4: Overall improvement for immediate lumbar imaging versus usual clinical care without immediate imaging

(Chou, 2009)

Another meta-analysis cited by the Institute for Clinical Systems Improvement guideline found that, "lumbar radiography and CT contribute to cumulative low-level radiation exposure, which could promote carcinogenesis. Lumbar spine CT is associated with an average effective radiation dose of 6 mSv (Fazel, 2009). On the basis of the 2.2 million lumbar CT scans performed in the United States in 2007, a study (Berrington de Gonzalez, 2009) projected 1,200 additional future cases of cancer" (Chou, 2011).

This meta-analysis also described that lumbar radiography occurs much more frequently than the CT scan, and therefore accounts for a greater proportion of the total radiation dose from lumbar imaging procedures in the United States (3.3% vs. 0.7%). The meta-analysis stated that, "the average radiation exposure from lumbar radiography is 75 times higher than for chest radiography (Fazel, 2009). This is of particular concern in young women because of the proximity to the gonads, which are difficult to effectively shield. The amount of female gonadal irradiation from lumbar radiography has been estimated as equivalent to having chest radiography daily for several years (Jarvik, 2002)" (Chou, 2011).

One meta-analysis identified psychosocial harm to the patient as another harm from unnecessary imaging of low back pain. For instance, telling a patient that they have a back imaging abnormality could result in unintended harms related to labeling, where a patient believes they have some type of malady when in fact they are healthy (Fisher, 1999). Any imaging study can sometimes produce clinically irrelevant results. For many patients, the knowledge of these findings might hinder recovery by causing increased worry and anxiety, excessive focus on minor back symptoms, and avoidance of exercise or other recommended activities due to fear of causing more structural damage (Fisher, 1999).

Low back imaging might also lead to unnecessary procedures. According to one study, "visual evidence can be very compelling, despite the uncertainties related to interpretation of most spinal imaging abnormalities, and imaging abnormalities may be viewed as targets for surgery or other interventions." (Rhodes, 1999). Another study found that for work-related acute low back pain, receiving an MRI within the first month was associated with more than an 8-fold increase in risk for surgery and more than a 5-fold increase in subsequent medical costs compared with patients who did not receive early MRI (Webster, 2010).

Citations:

Berrington de Gonza´lez A, Mahesh M, Kim KP, Bhargavan M, Lewis R, Mettler F, et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. Arch Intern Med. 2009; 169:2071-7. [PMID: 20008689]

Chou R, Fu R, Carrino JA, Deyo RA. 2009. "Imaging strategies for low-back pain: systematic review and meta-analysis." *The Lancet* 373(9662):463-72. (February 7, 2009) doi: 10.1016/S0140-6736(09)60172-0.

Chou R, Qaseem A, Owens DK, Shekelle P; Clinical Guidelines Committee of the American College of Physicians. 2011. "Diagnostic imaging for low back pain: advice for high-value health care from the American College of Physicians." *Annals of Internal Medicine* 154(3):181-9. (February 1, 2011) doi: 10.7326/0003-4819-154-3-201102010-00008.

Fazel R, Krumholz HM, Wang Y, Ross JS, Chen J, Ting HH, et al. Exposure to low-dose ionizing radiation from medical imaging procedures. N Engl J Med. 2009; 361:849-57. [PMID: 19710483]

Fisher ES, Welch HG. Avoiding the unintended consequences of growth in medical care: how might more be worse? JAMA. 1999; 281:446-53.

Institute for Clinical Systems Improvement (ICSI). Adult Acute and Subacute Low Back Pain. Updated November 2012. Guideline available from: https://www.icsi.org/_asset/bjvqrj/LBP.pdf, accessed February 6, 2014.

Jarvik JG, Deyo RA. Diagnostic evaluation of low back pain with emphasis on imaging. Ann Intern Med. 2002; 137:586-97. [PMID: 12353946]

Rhodes LA, McPhillips-Tangum CA, Markham C, Klenk R. The power of the visible: the meaning of diagnostic tests in chronic back pain. Soc Sci Med. 1999; 48:1189-203.

Webster BS, Cifuentes M. Relationship of early magnetic resonance imaging for work-related acute low back pain with disability and medical utilization outcomes. J Occup Environ Med. 2010; 52:900-7.

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

An important theoretical harm of this measure would be if patients who might benefit from early imaging do not receive imaging. We have supplied evidence that supports our supposition that the occurrence of such clinical events is quite rare.

There were very few harms described in the research regarding delaying lumbar imaging. A potential harm associated with delaying imaging is related to patient expectations; mentioned in a single study, patients assigned to receive routine imaging for uncomplicated low back pain were more likely to believe that the imaging was necessary, despite not experiencing any clinical benefit (Chou, 2009). In another trial (Kendrick, 2001), 80 percent of patients with low back pain would undergo radiography if given the choice, despite the lack of benefit from routine imaging. Since the harms related to imaging patients without red flags outweigh the benefits, the study concluded, "educational interventions could be effective for reducing the proportion of patients with low-back pain who believe that routine imaging should be done. We need to identify back-pain assessment and educational strategies that meet patient expectations and increase satisfaction, while avoiding unnecessary imaging" (Chou, 2009).

Citations:

Chou R, Deyo RA, Jarvik JG. 2012. "Appropriate use of lumbar imaging for evaluation of low back pain." *Radiologic Clinics of North America* 50(4):569-85. (July, 2012) doi: 10.1016/j.rcl.2012.04.005.

Chou R, Fu R, Carrino JA, Deyo RA. 2009. "Imaging strategies for low-back pain: systematic review and meta-analysis." *The Lancet* 373(9662):463-72. (February 7, 2009) doi: 10.1016/S0140-6736(09)60172-0.

Kendrick D, Fielding K, Bentley E, Kerslake R, Miller P, Pringle M. Radiography of the lumbar spine in primary care patients with low back pain: randomised controlled trial. BMJ 2001; 322: 400–05.

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

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

We are not aware of new evidence that would impact the current guideline on low back pain imaging.

1a.8 OTHER SOURCE OF EVIDENCE

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

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

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

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 0052
Corresponding Measures:
De.2. **Measure Title:** Use of Imaging Studies for Low Back Pain
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of patients with a primary diagnosis of low back pain who did not have an imaging study (plain X-ray, MRI, CT scan) within 28 days of diagnosis.
1b.1. **Developer Rationale:** This measure assesses the overuse of imaging studies (plain x-ray, MRI, and CT scans) in adults with acute, uncomplicated low back pain. The intent of this measure is to reduce inappropriate imaging for low back pain –
that is, imaging in the absence of "red flags" (indications that back pain is caused by a serious, underlying pathology that would warrant imaging). Inappropriate imaging is problematic because it is not associated with improved outcomes and exposes patients to unnecessary harms such as radiation exposure and further unnecessary treatment (Chou, Fu, Carrino and Deyo, 2009).

Chou R, Fu R, Carrino JA, Deyo RA. 2009. "Imaging strategies for low-back pain: systematic review and meta-analysis." Lancet 373:463-72.

S.4. **Numerator Statement:** Patients who received an imaging study (x-ray, CT, MRI) within the 28 days following a diagnosis of low back pain.
S.7. **Denominator Statement:** All patients 18 years as of January 1 of the measurement year to 50 years as of December 31 of the measurement year with a claim/encounter for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, with a principal diagnosis of low back pain during the Intake Period (January 1 – December 3 of the measurement year).
S.10. **Denominator Exclusions:** Because the intent of the measure is to assess imaging for patients with a new episode of low back pain, exclude patients with a recent diagnosis of low back pain.

Also, exclude any patient who had a diagnosis for which imaging is clinically appropriate. Any of the following meet criteria:

(1) Cancer
(2) Trauma
(3) Recent IV drug abuse
(4) Neurologic impairment
(5) HIV
(6) Spinal infection
(7) Major organ transplant
(8) Prolonged use of corticosteroids

De.1. **Measure Type:** Process
S.23. **Data Source:** Claims (Only)
S.26. **Level of Analysis:** Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Aug 10, 2009

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
FINAL_2016_Evidence_Form_0052_LBP.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure assesses the overuse of imaging studies (plain x-ray, MRI, and CT scans) in adults with acute, uncomplicated low back pain. The intent of this measure is to reduce inappropriate imaging for low back pain – that is, imaging in the absence of "red flags" (indications that back pain is caused by a serious, underlying pathology that would warrant imaging). Inappropriate imaging is problematic because it is not associated with improved outcomes and exposes patients to unnecessary harms such as radiation exposure and further unnecessary treatment (Chou, Fu, Carrino and Deyo, 2009).

Chou R, Fu R, Carrino JA, Deyo RA. 2009. "Imaging strategies for low-back pain: systematic review and meta-analysis." Lancet 373:463-72.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
UPDATED INFORMATION FOR AD-HOC REVIEW (2016)
The following data are from HEDIS data collection reflecting the most recent years of measurement prior to this measure reevaluation. Performance data is summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th, 90th percentile and interquartile range (IQR). Data is stratified by year, geographic region, and product line (i.e. commercial HMO and PPO combined, Medicaid HMO and PPO combined).

The following data demonstrate the variation in the rate of appropriate imaging for patients with low back pain across health plans. In 2014, there was a 15-point difference between plans in the 10th percentile and plans in

the 90th percentile for commercial plans and 15 points for Medicaid plans. These gaps in performance underscore the opportunity for improvement.

Commercial Rate (HMO and PPO Combined)
YEAR| MEAN | ST DEV | Min | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 75% | 6% | 54% | 68% | 71% | 75% | 80% | 83% | 9
2013 | 75% | 6% | 26% | 67% | 70% | 75% | 79% | 83% | 9
2012 | 75% | 6% | 56% | 67% | 70% | 75% | 79% | 82% | 9

Medicaid Rate (HMO and PPO Combined)
YEAR| MEAN | ST DEV | Min | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2014 | 75% | 6% | 55% | 68% | 71% | 75% | 78% | 83% | 7
2013 | 76% | 5% | 58% | 68% | 72% | 75% | 78% | 84% | 6
2012 | 76% | 6% | 58% | 68% | 72% | 75% | 79% | 82% | 8

These data come from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2014, HEDIS measures covered more than 171 million health plan members. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

Commercial (HMO and PPO Combined)
YEAR | N Plans | Average Denominator Size
2014 | 404 | 3593
2013 | 408 | 3677
2012 | 409 | 3964

Medicaid HMO
YEAR | N Plans | Average Denominator Size
2014 | 200 | 1264
2013 | 193 | 1054
2012 | 180 | 1123

The tables below highlight geographic variation in 2014 performance rates for both commercial and Medicaid plans. The average performance rates in top performing regions are six and five percentage points above the national average for commercial and Medicaid plans, respectively. Additionally, the performance rates in the top performing regions are 13 and 11 percentage points above the lowest performing regions for Commercial and Medicaid plans, respectively. This underscores opportunities for improvement among lower performing regions of the country.

2014 Commercial Rate (HMO and PPO Combined) by HHS Region
Region | N Plans | MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
Atlanta | 62 | 68% | 5% | 61% | 65% | 68% | 68% | 73% | 4
Boston | 44 | 78% | 6% | 71% | 74% | 79% | 83% | 85% | 9
Chicago | 80 | 76% | 5% | 71% | 72% | 75% | 79% | 83% | 7
Dallas | 40 | 71% | 6% | 64% | 68% | 70% | 75% | 78% | 7
Denver | 23 | 78% | 6% | 73% | 75% | 79% | 82% | 86% | 7
Kansas City | 36 | 77% | 5% | 74% | 75% | 77% | 79% | 81% | 4
New York | 34 | 76% | 5% | 71% | 73% | 75% | 79% | 81% | 6
Philadelphia | 54 | 74% | 5% | 69% | 71% | 74% | 77% | 79% | 6
San Francisco | 45 | 78% | 6% | 72% | 75% | 79% | 81% | 84% | 6
Seattle | 30 | 81% | 6% | 76% | 78% | 81% | 85% | 86% | 7

2014 Medicaid Rate (HMO and PPO Combined) by HHS Region

Region | N Plans | MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
Atlanta | 29 | 69% | 6% | 60% | 67% | 70% | 73% | 75% | 6
Boston | 13 | 75% | 3% | 71% | 73% | 75% | 77% | 78% | 4
Chicago | 42 | 76% | 5% | 69% | 72% | 76% | 79% | 82% | 7
Dallas | 17 | 74% | 2% | 71% | 73% | 74% | 75% | 77% | 2
Denver | 5 | 79% | 4% | 72% | 80% | 80% | 81% | 83% | 1
Kansas City | 9 | 74% | 4% | 69% | 70% | 74% | 77% | 77% | 7
New York | 11 | 75% | 3% | 72% | 73% | 75% | 77% | 78% | 4
Philadelphia | 27 | 74% | 6% | 70% | 73% | 77% | 83% | 92% | 10
San Francisco | 42 | 80% | 6% | 74% | 77% | 79% | 84% | 87% | 7
Seattle | 5 | 76% | 3% | 71% | 75% | 78% | 79% | 79% | 4

INFORMATION FROM PREVIOUS SUBMISSION (2014)
The following data are from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data is summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th, 90th percentile and interquartile range (IQR). Data is stratified by year and product line (i.e. commercial HMO and PPO combined, Medicaid HMO).

The following data demonstrate the variation in the rate of appropriate imaging for patients with low back pain across health plans. In 2012, there was a 15.5 point difference between plans in the 10th percentile and plans in the 90th percentile for commercial plans and 13.9 points for Medicaid plans. These gaps in performance underscore the opportunity for improvement.

Commercial Rate (HMO and PPO Combined)
YEAR| MEAN | ST DEV | Min | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2012 | 75% | 6% | 56% | 67% | 70% | 75% | 79% | 82% | 9
2011 | 74% | 6% | 45% | 66% | 69% | 74% | 79% | 82% | 9
2010 | 74% | 6% | 53% | 66% | 70% | 74% | 78% | 81% | 8

Medicaid Rate (HMO)
YEAR| MEAN | ST DEV | Min | 10TH | 25TH | 50TH | 75TH | 90TH | IQR
2012 | 76% | 6% | 58% | 68% | 72% | 75% | 79% | 82% | 8
2011 | 76% | 5% | 62% | 70% | 72% | 76% | 79% | 82% | 7
2010 | 75% | 6% | 58% | 67% | 72% | 76% | 80% | 82% | 8

These data come from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2012, HEDIS measures covered 107.3 million commercial health plan members and 21.7 million Medicaid HMO members. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

Commercial HMO
YEAR | N Plans | Median Denominator Size per plan
2012 | 210 | 835
2011 | 210 | 932
2010 | 232 | 1178

Commercial PPO
YEAR | N Plans | Median Denominator Size per plan
2012 | 199 | 2547
2011 | 189 | 2350
2010 | 171 | 2434

Medicaid HMO
YEAR | N Plans | Median Denominator Size per plan
2012 | 180 | 698
2011 | 162 | 749
2010 | 151 | 744

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
HEDIS data is stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). NCQA does not currently collect performance data stratified by race, ethnicity, or language. Escare et al. have described in detail the difficulty of collecting valid data on race, ethnicity and language at the health plan level (Escare, 2011). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data. These measures follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

Escare J.J., Carreon R., Vesolovskiy G., and Lawson E.H. 2011. Collection Of Race And Ethnicity Data By Health Plans Has Grown Substantially, But Opportunities Remain To Expand Efforts. Health Affairs 20(10): 1984-1991.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
We found limited data about disparities in the overuse of imaging for low back pain. One study from the Department of Veterans Affairs, using data from 110,661 outpatient MRIs of the lumbar spine, reported significantly higher rates of MRIs in younger adults (those under 35 years) compared to other ages (35-44 years, 45-54 years, 55-64 years, and older than 65; p<.0001). The study also reported significantly lower rates of MRIs in black adults compared to white adults (OR = 0.82, p<0.001) ((Gidwani R et al, 2016).

Gidwani R, Sinnott P, Avoundijian T, Lo J, Asch SM, Barnett PG. 2016. "Inappropriate ordering of lumbar spine magnetic resonance imaging: are providers Choosing Wisely?" Am J Manag Care 22(2): e68-76.

1c. High Priority (previously referred to as High Impact)
The measure addresses:
- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR

- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, High resource use
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Low back pain is a pervasive problem that affects three-quarters of adults at some time in their lives (Chou, 2012). Each year in the United States low back pain is experienced by 25 to 50 percent of adults, making it one of the most common reasons for seeking health care services (Haldeman, 2008). According to the U.S. Preventive Services Task Force, it is second only to upper respiratory problems as a symptom-related reason for visits to a physician (USPSTF, 2004), and accounts for over 4.7 million missed work days per year (Dagenis, 2008).

Low back pain also results in high indirect costs from disability, lost time from work, and decreased productivity while at work, and is the number one cause for activity limitations in younger adults (Chou, 2012). Given the high prevalence of back pain, it is not surprising that its economic consequences are severe. The costs associated with health care services for spine pain (primarily low back pain) in the U.S. increased from $45.9 billion in 1997 to $102.6 billion in 2004 (Martin, 2008). Research suggests that the reasons for the increase in cost and use of diagnostic imaging can be attributed to multiple factors including changing demographics, increased care seeking and patient expectations about low back pain, increased physician ownership of imaging facilities, and fee-for-service payment models (Pham, 2009). The supply of imaging equipment may also play a role, as the number of MRI scanners in the U.S. increased from 7.6 per 1 million people to 26.6 per 1 million people between 2000 and 2005 (Baras, 2009).

The three imaging modalities included in this measure are: x-ray, CT scan, and MRI, all of which have varying individual costs. Generally, the reimbursement rates and charges for lumbar spine CT run 5 to 10 times higher and MRI 10 to 15 times higher than low back radiography. Although radiography is relatively lower in cost, it represents a financial burden as it is much more frequently used than the two other imaging mechanisms. In 2004, an estimated 66 million lumbar radiographs were performed in the United States (Chou, 2012). These imaging practices directly affect the patient, and also result in downstream costs associated with invasive and expensive operations and procedures.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
Baras JD, Baker LC. Magnetic resonance imaging and low back pain care for Medicare patients. Health Aff 2009; 28:w1133–40.
Chou R, et al. Radiologic Clinics of North America. Appropriate Use of Lumbar Imaging for Evaluation of Low Back Pain. 2012 Jul, Vol. 50, No. 4: 569-85.
Dagenais S, et al. A systematic review of low back pain cost of illness studied in the United States and internationally. Spine Journal 2008; 8-1: 8-20
Dagenais S, et al, A systematic review of diagnostic imaging use for low back pain in the United States. Spine J 2013, doi: 10.1016/j.spinee.2013.10.031.
Haldeman S, Dagenais S. A supermarket approach to the evidence informed management of chronic low back pain. Spine J 2008;8: 1–7.
Martin BI, Deyo RA, Mirza SK, et al. Expenditures and health status among adults with back and neck problems. JAMA 2008;299: 656–64.
Pham HH, Landon BE, Reschovsky JD, et al. Rapidity and modality of imaging for acute low back pain in elderly patients. Arch Intern Med 2009; 169:972–81.

US Preventive Services Task Force. Primary Care Interventions to Prevent Low Back Pain: Brief Evidence Update (Feb 2004). Source: http://www.uspreventiveservicestaskforce.org/3rduspstf/lowback/lowbackup.htm.

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

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

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

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Musculoskeletal, Musculoskeletal : Low Back Pain

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

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 2016_0052_LBP_Value_Sets.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
In 2015, NCQA initiated a reevaluation of the Use of Imaging Studies for Low Back Pain measure used in the Healthcare Effectiveness Data and Information Set (HEDIS®). Changes to the measure were recommended based upon a review of evidence and guidelines, and feedback from numerous stakeholder groups and measurement advisory panels. The revisions were vetted through a 30-day public comment process and approved by NCQA's Committee on Performance Measurement (CPM). The intent of these changes was to better align the measure with the evidence and to improve the face validity of the measure. Both the changes and their rationale are detailed below.

(1) Include physical therapy and telehealth visits with a primary diagnosis of low back pain in the denominator.

Rationale: Harmonization with an existing measure. Under some health plans, individuals can self-refer to physical therapy, bypassing a physician visit. Including these visits could provide a more accurate index episode start date for a member's low back pain symptoms.

(2) Shorten the look-back period for the "recent trauma" exclusion from 12 months to 3 months.

Rationale: The longer 12-month look-back period may include past trauma that is unrelated to current low back pain complaints and inadvertently remove patients who should be assessed for inappropriate imaging by the measure.

(3) Exclude members with at least 90 consecutive days of corticosteroid use anytime in the past 12 months.

Rationale: Existing evidence indicates prolonged use of corticosteroids is significantly associated with fracture in individuals with low back pain; imaging may be appropriate in this scenario.

(4) Exclude members with HIV anytime in their history.

Rationale: Harmonization with an existing measure. Individuals with HIV are at increased risk for infection; imaging may be appropriate in this scenario.

(5) Exclude members with a major organ transplant anytime in their history.

Rationale: Individuals who have undergone a major organ transplant are at increased risk for infection due to continual treatment with immunosuppressive therapy; imaging may be appropriate in this scenario.

(6) Exclude members with current or recent (past 12 months) spinal infection (e.g., intraspinal abscess, osteomyelitis, discitis).

Rationale: Spinal infections are most often diagnosed through imaging. While most people with low back pain do not have a spinal infection, spinal infections often present with low back pain as a symptom.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who received an imaging study (x-ray, CT, MRI) within the 28 days following a diagnosis of low back pain.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who received an imaging study (see Imaging Study Value Set) with a diagnosis of low back pain (see Uncomplicated Low Back Pain Value Set) on the Index Episode Start Date (IESD) or in the 28 days following the IESD.

The Index Episode Start Date is the earliest date of service for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, during the Intake Period (January 1-December 3 of the measurement year) with a principal diagnosis of low back pain.

The measure is reported as an inverted rate (i.e. 1 – numerator/denominator). A higher score indicates appropriate treatment of low back pain (i.e. the proportion for whom imaging studies did not occur).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients 18 years as of January 1 of the measurement year to 50 years as of December 31 of the measurement year with a claim/encounter for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, with a principal diagnosis of low back pain during the Intake Period (January 1 – December 3 of the measurement year).

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

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
All patients 18 years as of January 1 of the measurement year to 50 years as of December 31 of the measurement year who had any of the following during the intake period (January 1 to December 3 of the measurement year):

(1) Outpatient visit (Outpatient Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).

(2) Observation visit (Observation Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set). Do not include observation visits that result in an inpatient stay (Inpatient Stay Value Set). An observation visit results in an inpatient stay when the ED/observation date of service and the admission date for the inpatient stay are one calendar day apart or less.

(3) ED visit (ED Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set). Do not include ED visits that result in an inpatient stay (Inpatient Stay Value Set). An ED visit results in an inpatient stay when the ED date of service and the admission date for the inpatient stay are one calendar day apart or less.

(4) Osteopathic or chiropractic manipulative treatment (Osteopathic and Chiropractic Manipulative Treatment Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).

(5) Physical Therapy visit (Physical Therapy Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).

(6) Telehealth visit (Telehealth Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Because the intent of the measure is to assess imaging for patients with a new episode of low back pain, exclude patients with a recent diagnosis of low back pain.

Also, exclude any patient who had a diagnosis for which imaging is clinically appropriate. Any of the following meet criteria:

(1) Cancer
(2) Trauma
(3) Recent IV drug abuse
(4) Neurologic impairment
(5) HIV
(6) Spinal infection
(7) Major organ transplant
(8) Prolonged use of corticosteroids

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Because the intent of the measure is to assess imaging for patients with a new episode of low back pain, exclude patients with a diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set) during the 180 days (6 months) prior to the IESD.

Also, exclude any patient who had a diagnosis for which imaging is clinically appropriate. Any of the following meet criteria:

(1) Cancer (Malignant Neoplasms Value Set, Other Neoplasms Value Set, History of Malignant Neoplasms Value Set) any time during the patient's history through 28 days after the IESD.

(2) Trauma (Trauma Value Set) any time during the 3 months (90 days) prior to the IESD through 28 days after the IESD.

(3) IV drug abuse (IV Drug Abuse Value Set) any time during the 12 months (1 year) prior to the IESD through 28 days after the IESD.

(4) Neurologic impairment (Neurologic Impairment Value Set) any time during the 12 months (1 year) prior to the IESD through 28 days after the IESD.

(5) HIV (HIV Value Set) any time during the patient's history through 28 days after the IESD.

(6) Spinal Infection (Spinal Infection Value Set) any time during the 12 months (1 year) prior to the IESD through 28 days after the IESD.

(7) Major organ transplant (Organ Transplant Other Than Kidney Value Set; Kidney Transplant Value Set) any time in the patient's history through 28 days after the IESD.

(8) Prolonged use of corticosteroids. 90 consecutive days of corticosteroid treatment any time during the 12 months (1 year) prior to and including the IESD.

To identify consecutive treatment days, identify calendar days covered by at least one dispensed corticosteroid (Table LBP-A). For overlapping prescriptions assume the patient started taking the second prescription after exhausting the first prescription. For example, if a patient had a 30-day prescription dispensed on June 1 and a 30-day prescription dispensed on June 26, there are 60 covered calendar days (June 1 – July 30).

Count only medications dispensed during the 12 months (1 year) prior to and including the IESD. When identifying consecutive treatment days, do not count days supply that extend beyond the IESD. For example, if a patient had a 90-day prescription dispensed on the IESD, there is one covered calendar day (the IESD).

No gaps are allowed.

Table LBP-A: Prescriptions to Identify Corticosteroids
Hydrocortisone; Cortisone; Prednisone; Prednisolone; Methylprednisolone; Triamcinolone; Dexamethasone; Betamethasone

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

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

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

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

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

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

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Identify all patients 18 years as of January 1 of the measurement year to 50 years as of December 31 of the measurement year who had any of the following visits during the Intake Period (i.e. January 1 – December 3):
• Outpatient visit (Outpatient Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).
• Observation visit (Observation Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set). Do not include observation visits that result in an inpatient stay (Inpatient Stay Value Set). An observation visit results in an inpatient stay when the ED/observation date of service and the admission date for the inpatient stay are one calendar day apart or less.
• ED visit (ED Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set). Do not include ED visits that result in an inpatient stay (Inpatient Stay Value Set). An ED visit

results in an inpatient stay when the ED date of service and the admission date for the inpatient stay are one calendar day apart or less.
• Osteopathic or chiropractic manipulative treatment (Osteopathic and Chiropractic Manipulative Treatment Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).
• Physical Therapy visit (Physical Therapy Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).
• Telehealth visit (Telehealth Value Set), with a principal diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set).

Step 2: Determine the Index Episode Start Date (IESD). The IESD is the earliest date of service for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, during the Intake Period (January 1-December 3 of the measurement year) with a principal diagnosis of low back pain. For each patient identified in step 1, determine the earliest episode of low back pain. If the member had more than one encounter, include only the first encounter.

Step 3: Exclude patients with a diagnosis of uncomplicated low back pain (Uncomplicated Low Back Pain Value Set) during the 180 days (6 months) prior to the IESD (i.e., test for Negative Diagnosis History).

Step 4: Exclude any patient who had a diagnosis for which imaging is clinically appropriate. Any of the following meet criteria:
• Cancer. Cancer any time during the patient's history through 28 days after the IESD. Any of the following meet criteria:
– Malignant Neoplasms Value Set.
– Other Neoplasms Value Set.
– History of Malignant Neoplasm Value Set.
• Recent trauma. Trauma (Trauma Value Set) any time during the 3 months (90 days) prior to the IESD through 28 days after the IESD.
• Intravenous drug abuse. IV drug abuse (IV Drug Abuse Value Set) any time during the 12 months (1 year) prior to the IESD through 28 days after the IESD.
• Neurologic impairment. Neurologic impairment (Neurologic Impairment Value Set) any time during the 12 months (1 year) prior to the IESD through 28 days after the IESD.
• HIV. HIV (HIV Value Set) any time during the patient's history through 28 days after the IESD.
• Spinal infection. Spinal Infection (Spinal Infection Value Set) any time during the 12 months (1 year) prior to the IESD through 28 days after the IESD.
• Major organ transplant. Major organ transplant (Organ Transplant Other Than Kidney Value Set; Kidney Transplant Value Set) any time in the patients's history through 28 days after the IESD.
• Prolonged use of corticosteroids. 90 consecutive days of corticosteroid treatment any time during the 12 months (1 year) prior to and including the IESD.

To identify consecutive treatment days, identify calendar days covered by at least one dispensed corticosteroid (Table LBP-A). For overlapping prescriptions assume the patient started taking the second prescription after exhausting the first prescription. For example, if a patient had a 30-day prescription dispensed on June 1 and a 30-day prescription dispensed on June 26, there are 60 covered calendar days (June 1 – July 30).

Count only medications dispensed during the 12 months (1 year) prior to and including the IESD. When identifying consecutive treatment days, do not count days supply that extend beyond the IESD. For example, if a patient had a 90-day prescription dispensed on the IESD, there is one covered calendar day (the IESD).

No gaps are allowed.

Table LBP-A: Prescriptions to Identify Corticosteroids
Hydrocortisone; Cortisone; Prednisone; Prednisolone; Methylprednisolone; Triamcinolone; Dexamethasone;

Betamethasone

Step 5: Calculate a rate (number of patients receiving an imaging study (i.e. plain x-ray, MRI, CT scan).

Step 6: Subtract the rate calculated in Step 6 from one to invert the measure result to represent appropriate treatment of low back pain (i.e. the proportion for whom imaging studies did not occur). The measure is reported as an inverted rate (i.e. 1- numerator/denominator) to reflect the number of people who did not receive an imaging study.

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

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

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

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

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

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

N/A
2a. Reliability – See attached Measure Testing Submission Form **2b. Validity – See attached Measure Testing Submission Form** FINAL_2016_Testing_Form_0052_LBP.docx

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.

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

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

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

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

AND

If patient preference (e.g., informed 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]

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

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

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*

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.
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

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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts

an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.
The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Health Plan Ranking http://reportcard.ncqa.org/plan/external/plansearch.aspx Annual State of Health Care Quality: http://www.ncqa.org/tabid/836/Default.aspx Physician Quality Reporting System (PQRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ Quality Rating System (Marketplace) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Health-Insurance-Marketplace-Quality-Initiatives.html Consensus Core Quality Measures Set: ACO and PCMH/Primary Care Measures https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/Core-Measures.html Health Plan Ranking http://reportcard.ncqa.org/plan/external/plansearch.aspx Annual State of Health Care Quality: http://www.ncqa.org/tabid/836/Default.aspx Physician Quality Reporting System (PQRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ Quality Rating System (Marketplace) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Health-Insurance-Marketplace-Quality-Initiatives.html Consensus Core Quality Measures Set: ACO and PCMH/Primary Care Measures https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/Core-Measures.html Payment Program California's Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p CMS Eligible Professional EHR Incentive Program (Meaningful Use) https://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives Regulatory and Accreditation Programs HEDIS Health Plan Accreditation http://www.ncqa.org/programs/accreditation/health-plan-hp HEDIS Accountable Care Organization (ACO) Core Performance Measures: http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx HEDIS Health Plan Accreditation http://www.ncqa.org/programs/accreditation/health-plan-hp HEDIS Accountable Care Organization (ACO) Core Performance Measures:

	http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx Quality Improvement (external benchmarking to organizations) Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Quality Compass http://www.ncqa.org/tabid/177/Default.aspx California's Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p Consensus Core Quality Measures Set: ACO and PCMH/Primary Care Measures https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/Core-Measures.html

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

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

ANNUAL STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2015 the report included data from 814 HMOs and 353 PPOs, representing more than 171 million patients.

CALIFORNIA'S VALUE BASED PAY FOR PERFORMANCE PROGRAM: This measure is used in the California P4P program, which is the largest non-governmental physician incentive program in the United States. Founded in 2001, it is managed by the Integrated Healthcare Association (IHA) on behalf of eight commercial HMO health plans representing 9 million insured persons. IHA reports results on approximately 35,000 physicians in 200 physician organizations.

CONSENSUS CORE QUALITY MEASURES SET: This measure is included in the ACO and PCMH / Primary Care Measure set within the Consensus Core Quality Measures Set. The Centers for Medicare & Medicaid Services (CMS), commercial plans, Medicare and Medicaid managed care plans, purchasers, physician and other care provider organizations, and consumers worked together through the Core Quality Measures Collaborative to identify core sets of quality measures that payers have committed to using for reporting as soon as feasible.

CMS ELIGIBLE PROFESSIONAL EHR INCENTIVE PROGRAM (MEANINGFUL USE): The Medicare and Medicaid Electronic Health Care Record (EHR) Incentive Programs provide incentive payments to eligible professionals as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology.

HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of commercial and Medicaid Health Plans. In 2012, a total of 77 Medicaid health plans were accredited using this measure covering 9.1 million members and 336 commercial health plans covering 87 million lives. Health plans are scored based on performance compared to benchmarks.

HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings which are reported in Consumer Reports and on the NCQA website. These ratings are based on performance on HEDIS measures among other factors. The 2015-2016 health plan ratings reviewed nearly 1,500 health plans and rated more than 1,000 private, Medicare and Medicaid health insurance plans.

HEDIS ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality,

reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

PHYSICIAN QUALITY REPORTING SYSTEM: This measure is used in the Physician Quality Reporting System (PQRS) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). Eligible professionals who satisfactorily report data on quality measures for covered Physician Fee Schedule services furnished to Medicare Part B beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer) receive these payment incentives and adjustments.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

QUALITY RATING SYSTEM: Quality Rating System (QRS) clinical measure data is submitted for Qualified Health Plans (QHP) as a condition of certification and participation in the Marketplaces. QRS data is used to create quality rating information in every Marketplace.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

From 2012 to 2014, the average performance rate of appropriate treatment of low back pain has remained around 75 percent for both commercial and Medicaid plans. Rates in the 10th and 90th percentile scores over the same years were approximately 67 percent and 83 percent for commercial plans and 68 percent and 84 percent for Medicaid plans, respectively. Data going back to 2005 reveals that the average performance scores have remained relatively unchanged, with averages ranging from 73-79 percent for both plan types from 2005-2014.

Although the performance rates have not improved, 2014 regional average performance rate data highlights a gap in performance rates across HHS regions for both commercial and Medicaid plans. The performance rates

for the plans in the lowest and highest performing regions is 69 percent and 80 percent for Medicaid plans and 68 percent and 81 percent for commercial plans, respectively.

We revised the measure in 2016 to include additional visit types in the denominator (i.e., telehealth and physical therapy visits) and to exclude individuals with indications of serious, underlying pathologies for low back pain (i.e., exclusions for HIV, spinal infection, major organ transplant, and prolonged use of corticosterioids). We made these revisions to better align the measure with the clinical guidelines; the changes may impact plan performance on this measure.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
The variation in scores between plans in the 10th percentile and 90th percentile, and the variation in regional average performance scores, indicate that plans with poorer performance can improve.

When considering changes to this measure, we sought feedback on adding additional exclusions that may affect the measure's performance. The addition of these exclusions could affect the measure's performance rate over time.

Increased public awareness of appropriate imaging for low back pain could also affect the measure's performance rate. Choosing Wisely, an initiative of the American Board of Internal Medicine Foundation in collaboration with more than 70 specialty society partners, promotes a, "national dialogue on avoiding wasteful or unnecessary medical tests, treatments and procedures" by publishing recommendations from the specialty societies to, "facilitate wise decisions about the most appropriate care based on a patient's individual situation." Since the release of the initial Choosing Wisely lists, six specialty societies have published recommendations regarding the use of imaging for patients with low back pain (Choosing Wisely, 2015), indicating the topic's importance to healthcare providers.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We will evaluate performance results in 2017, as well as feedback from stakeholders, to assess if the changes to the measure impacted performance.

5. Comparison to Related or Competing Measures

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

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

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0514 : MRI Lumbar Spine for Low Back Pain

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

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Concurrently with the most recent measure reevaluation, the NCQA and CMS measure teams compared the specifications for NQF #0052 and NQF #0514, and identified several opportunities for harmonization. NCQA and CMS measure teams shared this memo with NQF staff that describes several areas of harmonization between the two measures.

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
The NCQA measure (NQF #0052) addresses a different target population than the CMS measure (NQF #0514), and as such the measures are not competing measures.

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, swift@ncqa.org, 202-955-5174-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

UPDATED INFORMATION FOR AD-HOC REVIEW (2016)
Bone and Joint Measurement Advisory Panel
Elizabeth Drye, MD, MS, Yale/Yale New Haven Hospital
Cissy Kraft, MD, MS, MHS, FAAFP, Anthem BCBS of Colorado/Nevada
Kathy Lester, Morpace Inc.
Carolyn Oddo, PT, MS, FACHE, American Physical Therapy Association, Board of Directors
Jeffrey Susman, MD, Northeast Ohio Medical University

The NCQA Bone and Joint Measurement Advisory Panel advised NCQA during measure reevaluation. They evaluated the measure specification, reviewed field test results, and assessed NCQA's overall desirable attributes of Relevance, Scientific Soundness and Feasibility. The advisory panel consisted of a balanced group of experts, including representation from primary care. In addition to this advisory panel, we vetted the measure with a host of other stakeholders (see below). Thus, our measures are the result of consensus from a broad and diverse group of stakeholders.

Committee on Performance Measurement
Bruce Bagley, MD, American Medical Association & American Association for Physician Leadership
Andrew Baskin, MD, Aetna
Patrick Conway, MD, MSC, Center for Medicare & Medicaid Services
Jonathan D. Darer, MD, MPH, Medicalis
Helen Darling, Interim – National Quality Forum
Rebekah Gee, MD, MPH, FACOG, LSU School of Medicine and Public Health
Foster Gesten, MD, NY State Department of Health
David Grossman, MD, MPH, Group Health Physician
Christine S. Hunter, MD (Co-Chair)
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Nancy Lane, PhD, Vanderbilt University Medical Center
Bernadette Loftus, MD, The Permanente Medical Group
Amanda Parsons, MD, Montefiore Health System
J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, PhD, RN, AARP Public Policy Institute
Eric C Schneider, MD, MSc, FACP (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Blue Shield of California
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms

HEDIS Expert Coding Panel
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CHCA, CHDA, CCS, CCS-P, CDIP, CHTS-CP, CPHQ, Aqurate Health Data Management, Inc.
Elonia Griffin, RN, BSN, CareSource
Nelly Leon-Chisen, RHIA, American Hospital Association
Tammy Marshall, LVN, Aetna
Alec McLure, MPH, RHIA, CCS-P, Verisk Health
Michele Mouradian, RN, BSN, McKesson Corporation
Craig Thacker, RN, Cigna
Mary Jane Toomey, RN, CPC, WellCare Health Plans, Inc.

HEDIS Expert Pharmacy Panel
Linda DeLaet, PharmD, Kaiser Permanente
Gerry Hobson, RPh, Cerner Multum
Chronis H. Manolis, RPh, UPMC Health Plan
Cathrine Misquitta, PharmD, BCPS, FCSHP, Health Net Pharmaceutical Services
Kevin Park, MD, Molina Healthcare, Inc.

Technical Measurement Advisory Panel
Andy Amster, MSPH, Kaiser Permanente
Kathryn Coltin, MPH, Independent Consultant
Lekisha Daniel-Robinson, MSPH, Centers for Medicare and Medicaid Services
Marissa M. Finn, MBA, Cigna HealthCare
Scott Fox, MS, Med, Independence Blue Cross
Carlos Hernandez, CenCal Health
Kelly Isom, RN, MA, Aetna
Harmon S. Jordan, ScD, RTI International
Ernest Moy, MD, MPH, Agency for Healthcare Research and Quality (AHRQ)
Patrick Roohan, New York State Department of Health (NYSDOH)
Lynne Rothney-Kozlak, MPH, Rothney-Kozlak Consulting, LLC
Natan Szapiro, Independence Blue Cross

INFORMATION FROM PREVIOUS SUBMISSION (2014)
Musculoskeletal Work Group
Teresa Brady, PhD, Centers for Disease Control and Prevention
Saleh Khaled, MD MSc FRCSC, SIU- Division of Orthopedic Surgery
John Klippel, MD, Arthritis Foundation
Elizabeth Kraft, MD MHS FAAFP, Anthem Blue Cross Blue Shield in Colorado/Nevada
Catherine MacLean, MD, PhD, Greater Los Angeles Healthcare System
Tom Marr, MD, Health Partners
John Mason, PhD, BCBS of Massachusetts
Ken Saag, MD MSc, University of Alabama at Birmingham
Neil Wenger, MD, UCLA School of Medicine
Patience White, MD, Arthritis Foundation

Bone Joint Measure Advisory Panel
Elizabeth Kraft, MD MHS FAAP, Anthem Blue Cross Blue Shield, Colorado
Patience H. White, MD, National Arthritis Foundation
Thomas J. Marr, MD, HealthPartners
David Borenstein, MD, Arthritis and Rheumatism Associates
Kenneth Saag, MD, MSc, University of Alabama at Birmingham
Gunnar B. J. Andersson, M.D., Ph.D., Midwest Orthopaedics at Rush
Stephen C. Schoenbaum, MD, MPH, Josiah Macy Jr. Foundation
Ted Mikuls, MD, MSPH, Univ. of Nebraska / Omaha VAMC
Sarah Sampsel, MPH, WellPoint, Inc
Neil Wenger, MD, MPH, UCLA Department of Medicine

Committee on Performance Measurement (CPM)
Peter Bach, MD, Memorial Sloan Kettering Cancer Center
Bruce Bagley, MD, American Academy of Family Physicians
Andrew Baskin, MD, Aetna
A. John Blair lll, MD, Taconic IPA, Inc
Patrick Conway, MD, MSC, Center for Medicare & Medicaid Services
Jonathan D. Darer, MD, Geisinger Health System

Helen Darling, National Business Group on Health
Foster Gesten, MD, NYSDOH Office of Managed Care
Marge Ginsburg, Center for Healthcare Decisions
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management
George J. Isham, MD, MS, HealthPartners
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Arthur Levin, MPH, Center for Medical Consumers
Philip Madvig, MD, The Permanente Medical Group
J. Brent Pawlecki, MD MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, RN, PhD, AARP
Eric C. Schneider, MD, MSc (Co-Chair), RAND Corporation
Marcus Thygeson, MD, MPH Blue Shield of Califorina

Technical Measurement Advisory Panel
Andy Amster, MSPH, Kaiser Permanente
Kathryn Coltin, MPH, Harvard Pilgrim Health Care
Lekisha Daniel-Robinson, Centers for Medicare and Medicaid Services (CMS)
Marissa Finn, MBA, Cigna HealthCare
Scott Fox, MS MEd, AmeriHealth Caritas
Carlos Hernandez, CenCal Health
Kelly Isom, MA RN, Aetna
Harmon Jordan, ScD, RTI International
Ernest Moy, MD MPH, Agency for Healthcare Research and Quality
Patrick Roohan, NYSDOH Office of Health Insurance Programs
Lynne Rothney-Kozlak, MPH, Rothney-Kozlak Consulting, LLC
Natan Szapiro, Independence Blue Cross

HEDIS Expert Coding Panel
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Health Services Advisory Group
Nelly Leon-Chisen, RHIA, American Hospital Association
Tammy Marshall, LVN, Aetna
Alec McLure, RHIA, CCS-P, Verisk Health
Michele Mouradian, RN, BSN, McKesson Health Solutions
Craig Thacker, RN, CIGNA HealthCare
Mary Jane F. Toomey, RN CPC, Aetna Better Health

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2004
Ad.3 **Month and Year of most recent revision:** 07, 2016
Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance or otherwise modify the measures, and shall not disassemble, recompile or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0514

Measure Title: MRI Lumbar Spine for Low Back Pain

Measure Steward: Centers for Medicare & Medicaid Services

Brief Description of Measure: This measure evaluates the percentage of magnetic resonance imaging (MRI) of the lumbar spine studies for low back pain performed in the outpatient setting where conservative therapy was not attempted prior to the MRI. Antecedent conservative therapy may include claim(s) for physical therapy in the 60 days preceding the lumbar spine MRI, claim(s) for chiropractic evaluation and manipulative treatment in the 60 days preceding the lumbar spine MRI, or claim(s) for evaluation and management at least 28 days but no later than 60 days preceding the lumbar spine MRI. The measure is calculated based on a one-year window of Medicare claims data. The measure has been publicly reported, annually, by the measure steward, the Centers for Medicare & Medicaid Services (CMS), since 2010, as a component of its Hospital Outpatient Quality Reporting (HOQR) Program.

Developer Rationale: This measure will reduce overuse of imaging for uncomplicated low back pain without prior attempts at antecedent conservative therapy, as overuse in this population can result in detection of incidental findings and reflect poor care coordination. The measure score will guide patient selection of providers, assess quality, and inform quality improvement.

Numerator Statement: MRI of the lumbar spine studies with a diagnosis of low back pain (from the denominator) without the patient having claims-based evidence of prior antecedent conservative therapy.

Denominator Statement: The number of MRI of the lumbar spine studies with a diagnosis of low back pain on the imaging claim performed in a hospital outpatient department on Medicare FFS beneficiaries within a 12-month time window.

Denominator Exclusions: Below, in Section S.11 we provide a detailed list of denominator exclusion conditions. Denominator exclusions are consistent with current guidelines, evidence in literature, and guidance from the measure TEP.

Measure Type: Process

Data Source: Claims (Only)

Level of Analysis: Facility, Population : Regional and State

Original Endorsement Date: Oct 28, 2008 **Most Recent Endorsement Date:** Oct 28, 2008

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Summary of prior review in 2014

In the prior review, the evidence provided by the developer included 2007 American College of Radiology (ACR) Appropriateness Criteria low back pain (LBP) which recommends that uncomplicated acute LBP is a benign, self-limited condition that warrants no imaging studies. The 2007 ACR Appropriateness Criteria is based on a systematic review of 48 studies. Forty of the studies were rated category 3 and 4, with 4 being the lowest quality. None the studies were rated as category 1. In addition to the 2007 ACR Appropriateness Criteria, the total measure evidence provided included 14 additional guidelines.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided updated 2015 American College of Radiology (ACR) Appropriateness Criteria: Low Back Pain, based on a systematic review of 12 studies (3 well designed studies, 2 good quality studies, and 7 quality studies that may have design limitations); 18 supporting references also were included in the review. Six clinical variants are described. Procedures for uncomplicated acute low back pain (LBP) and/or radiculopathy with no "red flags" were rated as 1 or 2 (usually not appropriate), meaning that "the imaging procedure or treatment is unlikely to be indicated in the specified clinical scenarios, or the risk-benefit ratio for patients is likely to be unfavorable". Ratings for procedures used with other variants varied.

Exception to evidence: N/A

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review and grading conducted (Box 3) → QQC present (Box 4) → Evidence graded as moderate quality and SRs agree with not recommending imaging for non-specific low back pain→ Moderate

Questions for the Committee:

- *The evidence provided by the developer is updated but directionally the same as that presented for the previous NQF review. Does the Committee agree and so there is no need for repeat discussion and vote on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

The highest possible rating is MODERATE for evidence.

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

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

The developer provided an analysis of Medicare fee-for-service (FFS) claims data that indicates variation in the use of inappropriate MRI lumbar spine studies.

2011*	2012*	2013*	2014**	2015**	2016**
Measurement Period	Jan 2009–Dec 2009	Jan 2010–Dec 2010	Jan 2011–Dec 2011	Jul 2012–Jun 2013	Jul 2013–Jun 2014
Jul 2014–Jun 2015					
Facilities	1,128	1,128	1,128	1,128	1,128
Minimum Value	17.9%	12.3%	17.1%	17.6%	21.8%
14.9%					
5th Percentile	23.4%	26.7%	27.0%	28.0%	30.4%
29.1%					
25th Percentile	28.4%	32.0%	32.1%	33.1%	35.9%
35.3%					
Median	31.9%	35.8%	35.9%	36.8%	39.8%
39.0%					
75th Percentile	35.9%	40.1%	39.8%	41.0%	44.4%
43.5%					
95th Percentile	43.5%	48.6%	48.5%	48.0%	51.8%
50.6%					
Maximum Value		63.5%	69.1%	67.6%	67.7%
72.5%	64.8%				
Mean Performance (Standard Deviation)				32.5% (6.2)	
36.4% (6.8)	36.5% (6.6)	37.2% (6.2)	40.3% (6.6)	39.5% (6.6)	

Disparities
The developer used 2013 performance data to evaluate the effect of patient and facility characteristics on the likelihood of each beneficiary having an inappropriate MRI lumbar spine study. A logistic regression model was used to assess the relationship between patient and facility characteristics for the 207,573 MRI lumbar spine studies performed in 2013. The developer provided a summary of the data that showed beneficiary age, gender, race and facility characteristics had a substantial association with the rate of inappropriate MRI lumbar spine studies.
A summary of literature that highlights disparities associated with the overuse of imaging for low back between genders, racial groups and age bands was also provided.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Does this measure provide information to understand disparities in this area of healthcare?*

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

Committee pre-evaluation comments

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

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures **– no change in emphasis – specifications should be evaluated the same as with new measures**

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

Data source(s): Claims

Specifications:

- The level of analysis is at the facility and the care setting is clinician office/clinic; hospital: emergency department, acute care facility, imaging facility; urgent care: ambulatory. A lower score indicates better quality.
- The numerator is the number of lumbar spine MRIs in patients with a diagnosis of low back pain (from the denominator) without the patient having claims-based evidence of prior antecedent conservative therapy.
- The denominator includes the number of lumbar spine MRIs in patients with a diagnosis of low back pain on the imaging claim performed in a hospital outpatient department on Medicare FFS beneficiaries within a 12-month time window.
- The denominator exclusions include:
    - Patients with lumbar spine surgery in the 90 days prior to MRI
    - Cancer (within twelve months prior to MRI procedure)
    - Congenital spine and spinal cord malformations (within five years prior to MRI procedure)
    - Inflammatory and autoimmune disorders (within five years prior to MRI procedure)
    - Infectious conditions (within one year prior to MRI procedure)
    - Spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage (within five years prior to MRI procedure)
    - Spinal cord infarction (within one year prior to MRI procedure)
    - Neoplastic abnormalities (within five years prior to MRI procedure)
    - Treatment fields for radiation therapy (within five years prior to MRI procedure)
    - Spinal abnormalities associated with scoliosis (within five years prior to MRI procedure)
    - Syringohydromyelia (within five years prior to MRI procedure)
    - Postoperative fluid collections and soft tissue changes (within one year prior to MRI procedure)
    - Trauma (within 45 days prior to MRI procedure)
    - IV drug abuse (within twelve months prior to MRI procedure)
    - Neurologic impairment: (within twelve months prior to MRI procedure)
    - HIV (within twelve months prior to MRI procedure)
    - Unspecified immune deficiencies (within twelve months prior to MRI procedure)
    - Intraspinal abscess (an exclusion diagnosis must be in one of the diagnoses fields on the MRI lumbar spine claim)
- No updates have been made to the specification since the last evaluation in 2014.
- An attached spreadsheet contains numerous ICD-9 and ICD-10 codes for the diagnosis of low back pain. The diagnosis of low back pain must be on the MRI lumbar-spine claim to be included in the denominator.
- A calculation algorithm describes the process of calculating the measure and appears straightforward.
- This measure is not risk-adjusted.
- In the validity section, the developer discusses minimum case counts for the measure, but it is unclear whether this measure is specified only for those facilities that meet this requirement.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*
- *Are the exclusions listed appropriate for this measure?*
- *Is a minimum case count required for this measure?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:
For the prior review, empirical reliability testing was conducted using a signal-to–noise analysis with data from 100% Medicare Fee-For-Service Outpatient Standard Analytical Files (SAFs) from 2007-2011. The Committee was satisfied with the developer's interpretation of the measure score reliability testing.

Describe any updates to testing: The developer conducted new empirical testing at the measure score level that presented similar results to those presented in the prior review. Updates are described in RED.

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

Reliability was assessed using data obtained from 2013 Medicare FFS data. The testing sample was 1,616 facilities that met minimum case count requirements in 2013. A beta-binomial method was used to determine the ratio of signal to noise. A signal-to-noise analysis quantifies the amount of variation in a performance measure that is due to true differences (i.e., signal) as opposed to random measurement error (i.e., noise). Results will vary based on the amount of variation between the providers and the number of patients treated by each provider. This method results in a reliability statistic that ranges from 0 to 1. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences. A value of 0.7 often is regarded as a minimum acceptable reliability value. This is considered an appropriate test for measure score reliability.

Results of reliability testing: Reliability scores ranged from 22.4% to 86.6%, with a median reliability score of 44.9%.

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm: Precise specifications (Box 1) → empirical reliability testing (Box 2) → computed performance measure scores (Box 4) → signal-to-noise analysis (Box 5) → statistic and scope of testing (Box 6b) → Low.

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

RATIONALE: Although NQF does not require a particular threshold value for reliability estimates, a value of 0.7 often is used as a general rule-of-thumb. The higher the reliability value, the more confident one can be in the ability to distinguish the performance of one hospital from another. Lower values of the signal-to-noise reliability estimate may be the result of relatively little difference in hospital performance, relatively high levels of measurement error, relatively low sample size, or some combination. Additional analysis might shed light on why the reliability is low in the dataset tested (e.g., too few cases, which might suggest use of measure in larger facilities only).

The highest possible rating is HIGH for reliability, as testing was conducted at the performance score level.

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

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**
Specification not completely consistent with evidence: Exceptions somewhat align with "red flags" noted in the ACR Appropriateness Criteria.

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: Face validity was conducted by the developer.

Describe any updates to validity testing: Updated face validity testing results were included.

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

Method of validity testing of the measure score:
- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Previous validity testing method: Patient-level data and measure specifications were sent to 3,680 facilities in the HOQR program 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.

Updated validity testing method: Face validity of the measure score was systematically assessed through survey of a 11-member Technical Expert Panel (TEP). TEP members responded to the following questions:

1. Does NQF #0514 capture the most appropriate and prevalent types of antecedent conservative therapy available through claims data?
   (Response options: *yes*, *not sure* or *do not know*, and *no*)
2. The measure helps assess the inappropriate use of MRI lumbar-spine tests. Do you agree?
   (Response options: *strongly agree*, *agree*, *undecided*, *disagree*, *strongly disagree*, and *do not know*)

Note: 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.

Previous validity testing results: The developer reported that 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".*

Updated validity testing results:
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

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

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
  - *It is not clear whether the developer's question regarding ability of the measure to "help assess the inappropriate use of MRI lumbar-spine tests" explicitly address whether measure results can distinguish good from poor quality of care. Do you agree that it does and that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:
In the prior review, the Committed noted several concerns regarding exclusions, including:

- The interpretation of guidelines used in establishing exclusions for patients over 70 years of age. In some cases, the guidelines cited were in direct conflict with the measure exclusions. Other conflicts noted included suspected lumbar disc herniation, sciatica, acute radicular pain, spinal cord infarction or degenerative conditions.
- Concerns about the surgery exclusion look-back period, presentation of several conflicting guidelines used to identify 'red flag' conditions, and use of claims data to identify antecedent therapy.
- History of prior back surgery and previous trauma. The Committee noted that history of surgery should be an absolute exclusion, rather than a 90-day exclusion, as post-op back surgery patients cannot be categorized as uncomplicated back pain patients.

An updated analysis of measure exclusions was provided to determine the prevalence of each exclusion, by measured entity, and at an aggregate level. The developer 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 multiple categories of measure exclusions in 2013 performance data.

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

The frequency of excluded cases varied substantially across facilities (IQR: 12.61%). Median performance also changed substantially by applying the exclusion conditions. Median performance increases by 5.15% for facilities after applying measure exclusions. The developer stated that 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.

Questions for the Committee:

- *Did the developer address concerns regarding exceptions that were noted during the prior review?*
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

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

The developer states that risk adjustment or stratification is not appropriate for this process measure based on the measure evidence base and the measure construct.

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

- The developer noted that the performance results from measured entities that perform small numbers of MRI for low back pain could significantly skew results.
  - To address this, the developer applies minimum case count requirements before reporting performance scores.
  - The developer states the minimum case count requirements applied for this measure "assure a 95% confidence level for each provider's score."
- After applying the minimum case count restrictions, the developer reported a 95% confidence interval for each provider's score (n=1,616), noting that if a facility's score did not fall into this interval, it would be identified as better than or worse than average.
- Facility performance scores:

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

*Based on the reported mean and SD deviation, the 95% confidence interval is (24.09, 53.04) [mean +/- 1.96*SD → 38.57 +/- 14.48 → (24.09, 53.04)]. Fewer than 10 percent of facilities fell below this interval and fewer than 10% fell above this interval. The developer did not report how many of the 1,616 facilities had a performance value that was statistically significantly different from the mean.*

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*
- *Is the developer's approach to identify the minimum case count and the 95% confidence level appropriate? Does this mean that the measure is specified only for facilities that have a minimum case count?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The analytic files used for measure testing and measure calculation include post-adjudicated claims, and do not include known missing data.

Guidance from the Validity Algorithm Specifications somewhat with evidence (Box 1) → Assessed all potential threats to validity (Box 2) → no empirical testing (Box 3) →face validity assessed (Box 5) → Moderate, assuming potential threats to validity are not a problem or are adequately addressed.

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

Because only face validity was assessed rather than empirical validity, the highest rating possible rating is MODERATE for validity.

Committee pre-evaluation comments

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

Criterion 3. Feasibility

Maintenance measures **– no change in emphasis – implementation issues may be more prominent**

Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is claims-based, and uses CMS hospital outpatient claims as its data source. The data are coded by someone other than person obtaining original information.
- The developer notes that special attention needs to be taken when counting procedures on the Medicare claims files, specifically how to deal with modifier codes.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

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

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details: This measure is publically reported in the CMS Hospital Outpatient Quality Reporting program.

Improvement results:

- The developer reports that the median rate of overuse for facilities ***increased*** from 2011 to 2015 (31.9% to 39.0%).
- Performance has improved for subsets of facilities over the period of public reporting. For facilities that were classified as outliers in the first year of reporting (facilities with the highest 10% of performance scores, indicating poor performance) for the measure, only 16% were still outliers five years later. Facilities with persistent poor performance tend to be rural, small, and non-teaching.

Unexpected findings (positive or negative) during implementation: The developer states there were no unexpected findings during implementation.

Potential harms: The developer did not identify any unintended consequences during measure testing. No evidence of unintended consequences to individuals or populations have been reported to the developer since implementation.

Vetting of the measure: NQF has recently added a new subcriterion under Usability and Use: 4d: Vetting of the measure by those being measured and others is demonstrated when:

1) those being measured have been given performance results and data, as well as assistance with interpreting the measure results and data
2) those being measured and other users have been given an opportunity to provide feedback on the measure performance and implementation
3) this feedback has been considered when changes are incorporated into the measure

- The developer's submission does not include the items needed to evaluate whether vetting has been done for this measure. However, the developer notes a "dry-run" process that may meet NQF's requirements for vetting. The developer will be invited to provide additional information during the evaluation meeting.

Feedback :

No feedback provided on QPS. Measure reviewed by MAP for the Medicare Shared Savings Program in 2015. MAP conditionally supported this measure for MSSP pending resubmission to NQF for endorsement review.

Questions for the Committee*:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

How has the measure been vetted in real-world settings by those being measure or others?

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

Rationale: There appears to be little or no improvement in the measure results since 2011, with indications that results inappropriate MRIs for low back pain actually are increasing. Although the developer notes that earlier poor-

performing facilities have improved, it appears that performance among earlier higher-performing facilities may have worsened.

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

- **Competing measure:**0514: MRI Lumbar Spine for Low Back Pain (CMS)

Harmonization

- Due to differences in the level of analysis and care settings, the Committee will not be asked to select a best-in-class measure.
- Since the last evaluation, the developers have worked to harmonize the measures, resulting in greater congruence in how low back pain is defined, how cancer exclusions are defined, and in exclusion categories. Areas where the measures are not yet harmonized include the specific imaging modalities captured by the measure and some of the exclusion categories.

Endorsement + Designation

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

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

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

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because only a face validity assessment was conducted (rather than score-level empirical testing).

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

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

Measure Title: MRI Lumbar Spine for Low Back Pain

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

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

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

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

☒ Process: Overuse of magnetic resonance imaging (MRI) lumbar-spine studies for patients with low back pain for which there is no evidence of attempts at antecedent conservative therapy (2016 Submission).

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

☒ Other: Efficiency (2014 Submission)

1a.12 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

The process of identifying MRI studies of the lower back for patients for which antecedent conservative therapy has not yet been performed demonstrates instances of over usage. This awareness has led to incremental improved outcomes, , including attempts at non-invasive therapeutic procedures, better coordination of patient care, reduced exposure to contrast agents, and more efficient use of imaging resources.

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2014 Submission:** 1 - Bussieres AE, Taylor JA, Peterson C. Diagnostic imaging practice guidelines for musculoskeletal complaints in adults-an evidence-based approach-part 3: spinal disorders. J Manipulative Physiol Ther. 2008 Jan; 31(1):33–88. Guideline available at http://www.ncbi.nlm.nih.gov/pubmed/18308153. 2 - Chou R, Qaseem A, Snow V, et al. Clinical Efficacy Assessment Subcommittee of the American College of Physicians; American College of Physicians; American Pain Society Low Back Pain Guidelines Panel. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med. 2007; 147(7):478–91. Guideline available at http://www.ncbi.nlm.nih.gov/pubmed/17909209. 3 - Daffner RH, Wippold FJ II, Bennett DL, et al. Expert Panels on Musculoskeletal and Neurologic Imaging. ACR Appropriateness Criteria® suspected spine trauma. [online publication]. Reston (VA): American College of Radiology (ACR). 2012. Guideline available at http://www.guideline.gov/content.aspx?id=37931. 4 - Daffner RH, Weissman BN, Appel M, et al. Expert Panel on Musculoskeletal Imaging. ACR Appropriateness Criteria® stress (fatigue/insufficiency) fracture, including sacrum, excluding other vertebrae. [online publication]. Reston (VA): American College of Radiology (ACR). 2011. Guideline available at http://www.guideline.gov/content.aspx?id=32618. 5 - Davis PC, Wippold FJ II, Cornelius RS, et al. Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria® low back pain. [online publication]. Reston (VA): American College of Radiology (ACR). 2011. Guideline available at http://www.guideline.gov/content.aspx?id=35145. 6 – 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. Guideline available at http://www.guideline.gov/content.aspx?id=39319. 7 - Low back disorders. Occupational medicine practice guidelines: evaluation and management of common health problems and functional recovery in workers. 2nd ed. Elk Grove Village (IL): American College of Occupational and

	Environmental Medicine (ACOEM). 2007. Guideline available at http://www.guideline.gov/content.aspx?id=38438. 8 - Michigan Quality Improvement Consortium. Management of acute low back pain. Southfield (MI): Michigan Quality Improvement Consortium; 2012. Guideline available at http://www.guideline.gov/content.aspx?id=37956. 9 - Morrison WB, Zoga AC, Daffner RH, et al. Expert Panel on Musculoskeletal Imaging. ACR Appropriateness Criteria® primary bone tumors. [online publication]. Reston (VA): American College of Radiology (ACR). 2009. Guideline available at http://www.guideline.gov/content.aspx?id=15739. 10 - Practice Guideline for the Performance of MRI of the Adult Spine. Reston (VA): American College of Radiology (ACR). 2012. Guideline available at http://www.acr.org/~/media/ACR/Documents/PGTS/guidelines/MRI_Adult_Spine.pdf. 11 - Roberts CC, Daffner RH, Weissman BN, et al. Expert Panel on Musculoskeletal Imaging. ACR Appropriateness Criteria® metastatic bone disease. [online publication]. Reston (VA): American College of Radiology (ACR). 2012. Guideline available at http://www.guideline.gov/content.aspx?id=37930. 12 - Seidenwurm DJ, Wippold FJ II, Cornelius RS, et al. Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria® myelopathy. Reston (VA): American College of Radiology (ACR). 2011. Guideline available at http://www.guideline.gov/content.aspx?id=35146. 13 - University of Michigan Health System. Acute low back pain. Ann Arbor (MI): University of Michigan Health System. 2010 Jan. Guideline available at http://www.guideline.gov/content.aspx?id=23939. 14 - Wippold FJ II, Cornelius RS, Broderick DF, et al. Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria® dementia and movement disorders. [online publication]. Reston (VA): American College of Radiology (ACR). 2010. Guideline available at http://www.guideline.gov/content.aspx?id=32612. 15 - Work Loss Data Institute. Low back - lumbar & thoracic (acute & chronic). Corpus Christi (TX): Work Loss Data Institute. 2011. Guideline available at http://www.guideline.gov/content.aspx?id=33184. **2016 Submission:** • ACR Appropriateness Criteria® low back pain. • Patel ND, Broderick DF, Burns J, et al. Expert Panel on Neurologic Imaging. • 2015 • Patel ND, Broderick DF, Burns J, et al. Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria® low back pain. [online publication]. Reston (VA): American College of Radiology (ACR). 2015. 12 p. • https://acsearch.acr.org/docs/69483/Narrative.
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If	**2014 Submission:** Guideline # 1 – *Adult patient with acute uncomplicated* LBP (<4 weeks' duration.)* **Uncomplicated definition: non-traumatic LBP without neurologic deficits or indicators of potentially serious pathologies)—(see red flag list for details in the original guideline document).* *For most young or middle-aged adults, early diagnostic evaluation of low back complaints may focus on 3 basic questions (diagnostic imaging is infrequently required) (Jarvik, 2002).* *Is there underlying systemic disease?*

not a guideline, summarize the conclusions from the SR.	*Is there neurologic impairment that might require surgical intervention?* *Is social or psychological distress amplifying or prolonging the pain?* Radiographs not initially indicated [B] Special investigations not indicated [B] *Adult patient with uncomplicated subacute (4-12 wks.' duration) or persistent low back pain (LBP) (>12 wks.' duration) AND no previous treatment trial.* *A trial of up to 4-6 wk. of conservative care is appropriate before radiographs.* Radiographs not initially indicated [B] *Adult patient with non-traumatic acute LBP AND sciatica (no red flags).* *The first clinical clue to neurologic impairment usually is a history of sciatica: sharp pain radiating down the posterior or lateral aspect of the leg, often associated with numbness or paresthesia.* Radiographs not initially indicated [B] <u>Guideline # 2 –</u> Clinicians should conduct a focused history and physical examination to help place patients with low back pain into 1 of 3 broad categories: nonspecific low back pain, back pain potentially associated with radiculopathy or spinal stenosis, or back pain potentially associated with another specific spinal cause. The history should include assessment of psychosocial risk factors, which predict risk for chronic disabling back pain (strong recommendation, moderate-quality evidence). Clinicians should not routinely obtain imaging or other diagnostic tests in patients with nonspecific low back pain (strong recommendation, moderate-quality evidence). Clinicians should perform diagnostic imaging and testing for patients with low back pain when severe or progressive neurologic deficits are present or when serious underlying conditions are suspected on the basis of history and physical examination (strong recommendation, moderate-quality evidence). Clinicians should evaluate patients with persistent low back pain and signs or symptoms of radiculopathy or spinal stenosis with magnetic resonance imaging (preferred) or computed tomography only if they are potential candidates for surgery or epidural steroid injection (for suspected radiculopathy) (strong recommendation, moderate-quality evidence). Clinicians should provide patients with evidence-based information on low back pain with regard to their expected course, advise patients to remain active, and provide information about effective self-care options (strong recommendation, moderate-quality evidence). For patients with low back pain, clinicians should consider the use of medications with proven benefits in conjunction with back care information and self-care. Clinicians should assess severity of baseline pain and functional deficits, potential benefits, risks, and relative lack of long-term efficacy and safety data before initiating therapy (strong recommendation, moderate-quality evidence). For most patients, first-line medication options are acetaminophen or non-steroidal anti-inflammatory drugs. For patients who do not improve with self-care options, clinicians should consider the addition of non-pharmacologic therapy with proven benefits-for acute low back pain, spinal manipulation; for chronic or subacute low back pain, intensive interdisciplinary rehabilitation, exercise therapy, acupuncture, massage therapy,

	spinal manipulation, yoga, cognitive-behavioral therapy, or progressive relaxation (weak recommendation, moderate-quality evidence). Guideline # 3 – Clinical Condition: Suspected Spine Trauma **Variant 9:** Blunt trauma meeting criteria for thoracic or lumbar imaging. With or without localizing signs. MRI thoracic or lumbar spine without contrast: 5 MRI thoracic and lumbar spine without and with contrast: 1 **Variant 10:** Blunt trauma meeting criteria for thoracic or lumbar imaging. Neurologic abnormalities. MRI thoracic or lumbar spine without contrast: 9 MRI thoracic and lumbar spine without and with contrast: 1 Guideline # 4 – Clinical Condition: Stress (Fatigue/Insufficiency) Fracture, Including Sacrum, Excluding Other Vertebrae **Variant 1:** Suspect stress fracture. First imaging modality. MRI area of interest without contrast: 1 MRI area of interest without and with contrast: 1 **Variant 2:** Suspect stress fracture in patient with "need-to-know diagnosis", not hip or sacrum, Radiographs normal. MRI area of interest without contrast: 9 MRI area of interest without and with contrast: 1 **Variant 3:** Suspect stress fracture, not hip or sacrum. Radiographs normal. Bone scan positive and nonspecific. MRI area of interest without contrast: 9 MRI area of interest without and with contrast: 1 **Variant 4**: Suspect stress fracture in otherwise normal patient. Radiographs normal. MRI area of interest without contrast: 2 MRI area of interest without and with contrast: 1 **Variant 5:** Clinical differential fracture versus metastasis in long bone. Radiographs normal, bone scan hot but nonspecific. MRI area of interest without contrast: 9 MRI area of interest without and with contrast: 5 **Variant 6:** Clinical differential insufficiency fracture versus metastasis in sacrum. Radiographs normal, bone scan hot but nonspecific. MRI sacrum without contrast: 6 MRI sacrum without and with contrast: 1 **Variant 7**: Suspect insufficiency fracture in sacrum/pelvis; elderly patient. Radiographs normal. Bone scan hot in linear pattern typical for fracture. MRI pelvis without contrast: 6 MRI pelvis without and with contrast: 1 **Variant 8**: Suspect insufficiency fracture (any location) in osteoporotic patient or patient on long-term corticosteroid therapy. Radiographs normal. MRI area of interest without contrast: 9 MRI area of interest without and with contrast: 1 **Variant 9:** Suspect insufficiency fracture in osteoporotic patient or patient on long-term corticosteroid therapy. Radiographs and bone scan obtained within the preceding 48 hours are normal. MRI area of interest without contrast: 9 MRI area of interest without and with contrast: 1 Guideline # 5 – Clinical Condition: Low Back Pain

	Variant 1: Uncomplicated acute low back pain and/or radiculopathy, nonsurgical presentation. No red flags. (Red flags defined in the [original guideline]) MRI lumbar spine without contrast: 2 MRI lumbar spine without and with contrast: 2 **Variant 2:** Patient with one or more of the following: low velocity trauma, osteoporosis, focal and/or progressive deficit, prolonged symptom duration, age >70 years. MRI lumbar spine without contrast: 8 MRI lumbar spine with and without contrast: 3 **Variant 3:** Patient with one or more of the following: suspicion of cancer, infection, and/or immunosuppression. MRI lumbar spine without contrast: 7 MRI lumbar spine without and with contrast: 8 **Variant 4:** Low back pain and/or radiculopathy. Surgery or intervention candidate. MRI lumbar spine without contrast: 8 MRI lumbar spine without and with contrast: 5 **Variant 5:** Prior lumbar surgery. MRI lumbar spine without contrast: 6 MRI lumbar spine without and with contrast: 8 **Variant 6:** Cauda equina syndrome, multifocal deficits or progressive deficit. MRI lumbar spine without contrast: 9 MRI lumbar spine without and with contrast: 8 <u>Guideline # 6 –</u> Initial Evaluation and Data Set • Clinicians should not recommend imaging (including computed tomography [CT], magnetic resonance imaging [MRI] and x-ray) for patients with non-specific low back pain [Strong Recommendation, Moderate Quality Evidence] Core Treatment Plan • Clinicians should not recommend imaging (including computed tomography [CT], magnetic resonance imaging [MRI], and x-ray) for patients with non-specific low back pain [Strong Recommendation, Moderate Quality Evidence]. No Imaging First Six Weeks with Radicular Pain; Use Core Treatment Plan • Clinicians should not recommend imaging (including CT, MRI or x-ray) for patients in the first six weeks of radicular pain [Strong Recommendation, Moderate Quality Evidence]. <u>Guideline # 7 –</u> MRI for patients with acute LBP during the first 6 weeks if they have demonstrated progressive neurologic deficit, cauda equina syndrome, significant trauma with no improvement in atypical symptoms, a history of neoplasia (cancer), or atypical presentation (e.g., clinical picture suggests multiple nerve root involvement) – Recommended, Insufficient Evidence (I) MRI is not recommended for acute radicular pain syndromes in the first 6 weeks unless they are severe and not trending towards improvement and both the patient and the surgeon are willing to consider prompt surgical treatment, assuming the MRI confirms ongoing nerve root compression. Repeat MRI without significant clinical deterioration in symptoms and/or signs is also not recommended. – Not Recommended, Evidence (C) MRI is recommended for patients with subacute or chronic radicular pain syndromes lasting at least 4 to 6 weeks in whom the symptoms are not trending towards improvement if both the patient and surgeon are considering prompt surgical treatment, assuming the MRI confirms ongoing nerve root compression. In

	cases where an epidural glucocorticosteroid injection is being considered for temporary relief of acute or subacute radiculopathy, MRI at 3 to 4 weeks (before the epidural steroid injection) may be reasonable. – Moderately Recommended, Evidence (B) MRI is recommended as an option for the evaluation of select chronic LBP patients in order to rule out concurrent pathology unrelated to injury. This option should not be considered before 3 months and only after other treatment modalities (including NSAIDs, aerobic exercise, other exercise, and considerations for manipulation and acupuncture) have failed. – Recommended, Insufficient Evidence (I) Standing or weight-bearing MRI for any back or radicular pain syndrome or condition – Not Recommended, Insufficient Evidence (I) <u>Guideline # 8 –</u> Patients with High Risk of Serious Pathology (Red Flags and High Index of Suspicion) Spinal fracture or compressions—plain lumbosacral (LS) spine X-ray [B]. After 10 days, if fracture still suspected or multiple sites of pain, consider either bone scan [C] or referral [D] before considering computed tomography (CT) or magnetic resonance imaging (MRI). <u>Guideline # 9 –</u> Clinical Condition: Primary Bone Tumors *Variant 1* Screening, first study: MRI area of interest without or with contrast: 1 *Variant 2* Persistent symptoms, but radiograph negative: MRI area of interest without or with contrast: 9 *Variant 3* Definitively benign on radiographs (excluding osteoid osteoma): MRI area of interest without or with contrast: 1 *Variant 4* Clinically suspected osteoid osteoma: MRI area of interest without or with contrast: 6 *Variant 5* Suspicious for malignant characteristics on radiograph: MRI area of interest without or with contrast: 9 <u>Guideline # 10 –</u> Indications for spine MRI include, but are not limited to, the evaluation of: congenital spine and spinal cord malformations; inflammatory/autoimmune disorders; demyelinating disease; multiple sclerosis (MS); acute disseminated encephalomyelitis (ADEM); acute inflammatory demyelinating polyradiculopathy (Guillian-Barre syndrome); connective tissue disorders (e.g., systemic lupus erythematosus); infectious conditions; spinal infection, including disk space infection, vertebral osteomyelitis, and epidural abscess; spinal cord infection including abscess; vascular disorders; spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage; spinal cord infarction degenerative conditions; degenerative disk disease and its sequelae in the lumbar, thoracic, and cervical spine; neurodegenerative disorders such as subacute combined degeneration, spinal muscular atrophy, amyotrophic lateral sclerosis; trauma; nature and extent of injury to spinal cord, vertebral column, ligaments, thecal sac, and paraspinal soft tissues following trauma; neoplastic abnormalities intramedullary tumors; intradural extramedullary masses; intradural leptomeningeal disease; extradural soft tissue and bony neoplasms; treatment fields for radiation therapy; miscellaneous spinal abnormalities associated with scoliosis; syringohydromyelia (multiple etiologies, including Chiari malformations, trauma, etc); postoperative fluid collections and soft tissue changes (extradural and intradural); and pre-procedure assessment for vertebroplasty and kyphoplasty. <u>Guideline # 11 –</u> Clinical Condition: Metastatic Bone Disease

	Variant 3 Breast Carcinoma. Follow-up bone scan reveals single "hot" lesion in spine: MRI spine without contrast: 9 | MRI spine without and with contrast: 1 *Variant 4* Breast carcinoma. Three "hot" areas in spine revealed by bone scan. No back pain: MRI spine without contrast: 9 | MRI spine without and with contrast: 1 *Variant 9* Patient with known malignancy, with back pain and partially collapsed vertebra on radiography. Otherwise healthy : MRI spine without contrast: 9 | MRI spine without and with contrast: 1 *Variant 11* Patient with multiple myeloma presenting with acute low back pain: MRI lumbar spine without contrast: 8 | MRI lumbar spine without and with contrast: 1 <u>Guideline # 12 –</u> Clinical Condition: Myelopathy *Variant 1* Traumatic: MRI spine without contrast: 8 | MRI spine without and with contrast: 2 | MRI spine flow without contrast: 2 *Variant 2* Painful: MRI spine without contrast: 8 | MRI spine without and with contrast: 7 | MRI spine flow without contrast: 2 *Variant 3* Sudden Onset: MRI spine without contrast: 9 | MRI spine without and with contrast: 8 | MRI spine flow without contrast: 2 *Variant 4* Stepwise Progressive: MRI spine without contrast: 9 | MRI spine without and with contrast: 9 | MRI spine flow without contrast: 2 *Variant 5* Slowly Progressive: MRI spine without contrast: 9 | MRI spine without and with contrast: 9 | MRI spine flow without contrast: 2 *Variant 6* Infectious Disease Patient: MRI spine without contrast: 8 | MRI spine without and with contrast: 9 | MRI spine flow without contrast: 2 *Variant 7* Oncology Patient: MRI spine without contrast: 9 | MRI spine without and with contrast: 8 | MRI spine flow without contrast: 2 <u>Guideline # 13 –</u> Initial Visit: • Assess for "red flags" of serious disease (see Table 1 in the original guideline document), as well as psychological and social risks for chronic disability (see Table 2 in original guideline document). Diagnostic tests are usually unnecessary [IC]. If a patient has a red flag, obtain magnetic resonance imaging (MRI) and refer to specialist as appropriate. • X-rays, MRI, or computed tomography (CT) scan are not recommended for routine evaluation of patients with acute low back problems within the first 4-6 weeks of symptoms unless a red flag and high index of suspicion is noted on clinical evaluation. • For radicular pain without weakness, by ≥3 weeks: If no improvement, obtain MRI [IIB]. If not diagnostic, obtain electromyography (EMG). If pathology proven, consider evaluation by specialist in back pain or surgical evaluation [IA]. If pathology not proven, consider referral to specialist in back pain [ID]. Although opioid pain medications are effective [IIA], they are generally not indicated as first-line treatment and early opioid use may be associated with longer disability controlling for case severity [IIC]. <u>Guideline # 14 –</u> Clinical Condition: Dementia and Movement Disorders *Variant 12* Motor neuron disease: MRI spine without contrast: 8 | MRI spine without and with contrast: 7 <u>Guideline #15 –</u> Identify Radicular Signs

	• First visit: may be with Primary Care Physician MD/DO (50%), Orthopedist (33%), or Chiropractor (17%) (or rarely other specialists, including pain specialists)
	• Determine presence or absence of radiculopathy:
	  ○ Medical history
	  ○ Sensation: Feeling pain radiating below the knee (calf or lower), not just referred pain (pain radiating to buttocks or thighs), and dermatological sensory loss
	  ○ Straight leg raising test (sitting and supine), productive of leg pain
	  ○ Motor strength and deep tendon reflexes
	  ○ Document flexibility/range of motion (ROM) (fingertip test), muscle atrophy (calf measurement), local areas of tenderness, visual pain analog, sensation alternation
	• Rule out "red flag" diagnoses, including diagnostic studies, for specialist referral:
	  ○ Cauda Equina Syndrome (Schedule emergency procedure) (Refer to the original guideline document for International Classification of Diseases, Ninth Revision [ICD-9] codes for this and other diagnoses)
	  ○ Fracture, Compression fracture, Dislocation, Wound
	  ○ Cancer, Infection
	  ○ Dissecting/Ruptured Aortic Aneurysm
	  ○ Others (prostate problems, endometriosis/gynecological disorders, urinary tract infections, and renal pathology)
	Without Radiculopathy (90% of cases)
	• Also first visit (day 1):
	  ○ Prescribe activity modification, if necessary, based on severity and difficulty of job, while encouraging return to activity as much as possible; limited passive therapy with heat/ice (3 to 4 times/day); stretching/exercise (training by physical therapist OK); appropriate analgesia (i.e., acetaminophen) and/or anti-inflammatory (i.e., ibuprofen) [Benchmark cost: $14]; back to work except for severe cases in 72 hours, possibly modified duty; AVOID bed rest.
	  ○ REASSURE PATIENT: Patient education - common problem (90% of patients recover spontaneously in 4 weeks)
	  ○ No x-rays unless significant trauma (e.g., a fall)
	  ○ If muscle spasms, then consider muscle relaxant with limited sedative side effects [Benchmark cost: $44]
	• Second visit (day 3 to 10 - about 1 week after first visit, or sooner, because delayed treatment is not recommended)
	  ○ Document progress (flexibility, areas of tenderness, motor strength, straight leg raise--sitting and supine)
	  ○ If still 50% disabled (i.e., cannot return to work) then consider referral for exercise/instruction/manual therapy [Benchmark cost: $250]: Options are physical therapist, chiropractor, massage therapist, or occupational therapist (3 visits in first week), or by treating DO/MD (Choose providers supporting active therapy and not just passive modalities. The focus of treatment should not be symptom reduction, but improving function with a

	goal to return to work.) Consider screening for psychosocial symptoms in cases with expectations of delayed recovery. o Discontinue muscle relaxant • Third visit (day 10 to 17 - about 1 week after second visit) o Document progress o Prescribe muscle-conditioning exercises o At this point 66% to 75% should be back to regular work o While not indicated in the absence of red flags, if still disabled, then consider imaging study (anterior-posterior [AP]/lateral 2-view x-ray of lumbar) [Benchmark cost: $150] to rule out tumor, fracture, osteoporosis, myelopathy o Maintain therapy, continue focus on active therapy and not passive modalities, 2 visits in next week, teach home exercises o End manual therapy at 4 weeks (1 visit in last week) With Radiculopathy (10% of cases) • Also first visit (day 1) o Same as non-radicular • Second visit (day 3 to 10 - about 1 week after first visit) o Same as non-radicular, but o Reassure, but if increased numbness or weakness of either leg, get back to provider in one day o Consider referral to nonsurgical musculoskeletal physician (Orthopedist/Physical Medicine/Sports Medicine) • Third visit (day 10 to 17 - about 1 week after second visit) o Same as non-radicular, but o About 50% can be back at modified duty o If improvement, then add strengthening exercises, increased activity • Fourth visit (day 21 to 28 - about 1 to 2 weeks after third visit) o Document objective findings, if no improvement then: o First magnetic resonance imaging (MRI) (about 3% of total cases, or 30% of radicular cases) to confirm extruded disk with nerve root displacement (≥1 month conservative therapy) [Benchmark cost: $1,600] o MRI or computed tomography (CT) not indicated without obvious clinical level of nerve root dysfunction, clear radicular findings, or before 3 to 4 weeks o EMGs (electromyography) may be useful to obtain unequivocal evidence of radiculopathy, after 4 to 8 weeks conservative therapy, but EMGs are not necessary if radiculopathy is already clinically obvious o Consider an epidural steroid injection (ESI) for severe cases hoping to avoid surgery [Benchmark cost: $676] o If psychological factors retarding recovery are suspected, possibly refer to psychologist for testing [Benchmark cost: $540]. o Education: Consider back school as an option, if available. o If no improvement 7 to 14 days after the first ESI, consider prescribing 2nd ESI [Benchmark cost: $615]; there should be a maximum of two ESIs, and the second ESI can be 7 to 14 days after the first, depending upon the patient's response and functional gain

- Surgery (three months or more -- after appropriate work-up and consultation, concordance between radicular findings on radiologic evaluation and physical exam findings) (about 2% of total cases, or 20% of radicular cases) (See also ODG Indications for Surgery™ -- Discectomy in Procedure Summary of the original guideline document). Unequivocal objective findings are required based on neurological examination and testing.
  - Refer to fellowship trained Spine Surgeon: Neurosurgeon (50%), Orthopedist (50%)
  - Before surgery, screen for psychological symptoms that could affect surgical outcome (e.g., substance abuse, child abuse, work conflicts, somatization, verbalizations, attorney involvement, smoking)
  - Review options/outcomes with patient, let patient be part of decision making
  - Simple discectomy/laminectomy, minimally invasive [Benchmark cost: $17,400]
  - Post-operative pain, walking exercises, physical therapy
- Failure to recover: See the Procedure Summary (in the original guideline document) for options that may be available, along with links to the medical evidence. Also, see the NGC summary of the Work Loss Data Institute's guideline Pain (chronic).

2016 Submission:

Variant Number	Variant of Clinical Condition	MRI lumbar spine without contrast	MRI lumbar spine without and with contrast
1	Acute, subacute, or chronic uncomplicated low back pain or radiculopathy. No red flags. No prior management.	2	2
2	Acute, subacute, or chronic uncomplicated low back pain or radiculopathy. One or more of the following: low velocity trauma, osteoporosis, elderly individual, or chronic steroid use.[1]	7	Not Rated
3	Acute, subacute, or chronic low back pain or radiculopathy. One or more of the following:	7	8

[1] While ACR lists elderly individuals as patients for whom an MRI lumbar spine is appropriate, the accompanying literature review developed by ACR notes, "[a] recent study found no statistically significant difference in primary outcomes after 1 year for older adults who had spine imaging within 6 weeks after an initial visit for care for low back pain versus similar patients who did not undergo early imaging; thus this panel does not include age older than 50 as an independent red flag." As NQF #0514 excludes patients with red-flag conditions, elderly patients are not removed from the measure, aligning with findings from the ACR literature review to support guideline development.

		suspicion of cancer, infection, or immunosuppression.		
	4	Acute, subacute, or chronic low back pain or radiculopathy. Surgery or intervention candidate with persistent or progressive symptoms during or following 6 weeks of conservative management.	8	5
	5	Low back pain or radiculopathy. New or progressing symptoms or clinical findings with history of prior lumbar surgery.	6	8
	6	Low back pain with suspected cauda equina syndrome or rapidly progressive neurologic deficit.	9	8

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade

2014 Submission:

<u>The following grading scale applies to recommendations from guideline #1:</u>

Grades of Recommendation: This tool has been developed to grade recommendations according to the strength of available scientific evidence (level A to D)

A: At least one meta-analysis, systematic review or RCT rated as 1++, and directly applicable to the target population; or a systematic review of RCTs or a body of evidence consisting principally of studies rated as 1+,directly applicable to the target population and demonstrating overall consistency of results

B: A body of evidence including studies rated as 2++, directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 1++ or 1+

C: A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 2++**

D: Evidence level 3 or 4; or extrapolated evidence from studies rated as 2+; or evidences from trials classified as (minus) regardless of the level /

<u>The following grading scale applies to recommendations from guideline #2:</u>

Quality of Evidence	Strength of Recommendation: Benefits Do or Do Not Clearly Outweigh Risks	Strength of Recommendation: Benefits and Risks and Burdens are Finely Balanced
High	Strong	Weak
Moderate	Strong	Weak
Low	Strong	Weak
Insufficient evidence to determine net benefits or harms	I Recommendation	

Method for grading the strength of the overall evidence for an intervention:

Good: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes (at least 2 consistent, higher-quality trials)

Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, size, or consistency of

	included studies; generalizability to routine practice; or indirect nature of the evidence on health outcomes (at least 1 higher-quality trial of sufficient sample size; 2 or more higher-quality trials with some inconsistency; at least 2 consistent, lower-quality trials, or multiple consistent observational studies with no significant methodological flaws). *Poor:* Evidence is sufficient to assess effects on health outcomes because of limited number or power of studies, large and unexplained inconsistency between higher-quality trials, important flaws in trial design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes. <u>The following grading scale applies to recommendations from guideline #'s 3, 4, 5, 9, 11, 12, and 14:</u> Rating Scale: 1,2,3 Usually not appropriate 4,5,6 May be appropriate 7,8,9 Usually appropriate <u>The following grading scale applies to recommendations from guideline #'s 6 and 7:</u> Strength of Evidence Ratings *A:* Strong evidence-base: Two or more high-quality studies.* *B:* Moderate evidence-base: At least one high-quality study or multiple moderate-quality studies** relevant to the topic and the working population. *C:* Limited evidence-base: At least one study of moderate quality. *I:* Insufficient evidence: Evidence is insufficient or irreconcilable. *For therapy and prevention, randomized controlled trials (RCTs) or crossover trials with narrow confidence intervals and minimal heterogeneity. For diagnosis and screening, cross sectional studies using independent gold standards. For prognosis, etiology or harms, prospective cohort studies with minimal heterogeneity. **For therapy and prevention, well-conducted cohort studies. For prognosis, etiology or harms, well-conducted retrospective cohort studies or untreated control arms of RCTs. <u>The following grading scale applies to recommendations from guideline #8:</u> Levels of Evidence for the Most Significant Recommendations *A:* Randomized controlled trials *B:* Controlled trials, no randomization *C:* Observational studies *D:* Opinion of expert panel <u>The following grading scale applies to recommendations from guideline #10:</u> The guideline does not report a grading scale. <u>The following grading scale applies to recommendations from guideline #13:</u> Levels of Evidence for the Most Significant Recommendations — Strength of Recommendation *A:* Randomized controlled trials — *I:* Generally should be performed *B:* Controlled trials, no randomization — *II:* May be reasonable to perform *C:* Observational trials — *III:* Generally should not be performed *D:* Opinion of expert panel <u>The following grading scale applies to recommendations from guideline #15:</u> *Ranking by Type of Evidence* 1. Systematic Review/Meta-Analysis 2. Controlled Trial - Randomized (RCT) or Controlled 3. Cohort Study-Prospective or Retrospective 4. Case Control Series 5. Unstructured Review

	6. Nationally Recognized Treatment Guideline (from www.guideline.gov) 7. State Treatment Guideline 8. Other Treatment Guideline 9. Textbook 10. Conference Proceedings/Presentation Slides *Ranking by Quality within Type of Evidence* 1. High Quality 2. Medium Quality 3. Low Quality **2016 Submission:** Patel et al. referenced 30 articles within the Appropriateness Criteria®. Three of the studies were assigned to category 1 (defined as *the study is well-designed and accounts for common biases*). Two of the studies were assigned to category 2 (defined as *the study is moderately well-designed and accounts for most common biases*). Seven of the studies were assigned to category 3 (defined as *there are important study design limitations*). The final 18 studies were assigned to category 4 (defined as t*he study is not useful as primary evidence; the article may not be a clinical study or the study design is invalid, or conclusions are based on expert consensus*). The guideline notes that "while there are references that report on studies with design limitations, 5 well designed or good quality studies provide good evidence."
Provide all other grades and definitions from the evidence grading system	**2014 Submission:** The full rating systems are provided in Section 1.a4.3. **2016 Submission:** Category 1 - The study is well-designed and accounts for common biases. Category 2 - The study is moderately well-designed and accounts for most common biases. Category 3 - There are important study design limitations. Category 4 - The study is not useful as primary evidence. The article may not be a clinical study or the study design is invalid, or conclusions are based on expert consensus.
Grade assigned to the **recommendation** with definition of the grade	**2014 Submission:** The systematic review that served as the basis for guideline development identified 48 relevant studies. Each study was rated based on the following quality scale: *Category 1* The study is well-designed and accounts for common biases. *Category 2* The study is moderately well-designed and accounts for most common biases. *Category 3* There are important study design limitations. *Category 4* The study is not useful as primary evidence. The article may not be a clinical study or the study design is invalid, or conclusions are based on expert consensus. For example: a) the study does not meet the criteria for or is not a hypothesis-based clinical study (e.g., a book chapter or case report or case series description); b) the study may synthesize and draw conclusions about several studies such as a literature review article or book chapter but is not primary evidence; c) the study is an expert opinion or consensus document.

	Eight of the studies were assigned to category 2. Nine of the studies were assigned to category 3. The final 31 studies were assigned to category 4. **2016 Submission:** Recommendations made within the ACR guideline ranged from a value of 2 (defined as *Usually Not Appropriate)* through 9 (defined as *Usually Appropriate*). The evidence supporting these recommendations demonstrates consensus within the clinical community that MRI of the lumbar spine is not appropriate for patients presenting with uncomplicated low back pain and that imaging should only be used when the low back pain is in conjunction with a red-flag condition or scenario (as defined by recommendations graded with values of 7 through 9, considered *usually appropriate)*.
Provide all other grades and definitions from the recommendation grading system	**2014 Submission:** All grades are described in Section 1a.7.2. **2016 Submission:** 1, 2, 3 Usually not appropriate 4, 5, 6 May be appropriate 7, 8, 9 Usually appropriate
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2014 Submission:** The body of evidence evaluated for clinical guideline #5 includes three experimental studies, fifteen observational studies, and thirty reviews or other study designs. Experimental studies ranged in size from 47 to 380 cases. Observational studies ranged in size from 20 to 736 cases. For reviews and other study designs, the sample size ranged from 1 to 474 patients. The quantity and quality of the body of evidence is further bolstered by the literature used for guideline development for the other guidelines that support this measure. Results cited in this body of evidence are consistent across studies and guidelines. **2016 Submission:** The body of evidence evaluated for the ACR guideline includes 2 experimental studies, 10 observational studies, and 18 reviews or other study designs. Experimental studies ranged in size from 246 to 380 cases. Observational studies ranged in size from 23 to 5,239 cases. The quantity and quality of the body of evidence is further bolstered by the literature used for guideline development for the other guidelines that support this measure.
Estimates of benefit and consistency across studies	**2014 Submission:** Given the high costs associated with performing unnecessary MRI lumbar spine studies and the fact that MRI lumbar spine studies are inappropriate prior to conservative therapy, the overall net benefit in reducing overuse of MRI lumbar spine studies is a reduction in cost and a reduction in the number of procedures performed, per beneficiary. **2016 Submission:**

	Given the high costs associated with performing unnecessary MRI lumbar spine studies and the fact that MRI lumbar spine studies are generally inappropriate prior to attempting conservative therapy, the net benefit in reducing overuse of MRI lumbar spine studies is a reduction in cost and a reduction in the number of downstream procedures performed, per beneficiary.
What harms were identified?	**2014 Submission:** No harms in measure implementation were identified to counter the net benefit of the measure. **2016 Submission:** No harms in measure implementation were identified to counter the net benefit of the measure.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2014 Submission:** In addition to the fifteen guidelines cited above, a review of the clinical literature was conducted during the measure contractor's annual review of the literature for additional evidence and/or new studies that substantiate the measure's intent. Citations and summaries for the 28 items included in this review can be found in Section 1a.8.2. Some of these 28 studies have been published since the period of guideline development. Results cited in these studies are consistent across studies and with the guidelines cited above. **2016 Submission:** During the measure contractors annual review of the literature, there were no newly identified articles that changed the conclusions presented in the systematic review used to create the ACR Appropriateness Criteria® for low back pain. A review of the clinical literature was conducted during the measure contractor's annual review of the literature for additional evidence and/or new studies that support the measure's intent. The measure contractor identified relevant peer-reviewed publications by searching the PubMed MEDLINE database from January 1, 2014 to January 15, 2016, limiting included results to those published in the English language and that had abstracts available in PubMed. This search initially identified 781 articles; a further review by the contractor's clinical and measure-development team resulted in the inclusion of 26 articles. All newly identified articles supported the current measure specifications. In addition to the red-flag conditions supported by the ACR guideline, the current measure specifications further exclude 10 red-flag conditions (i.e., lumbar spine surgery 90 days prior to MRI, congenital spine and spinal cord malformations, spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage, spinal cord infarction, treatment fields for radiation therapy, spinal abnormalities associated with scoliosis, syringohydromyelia, post-operative fluid and soft tissue changes, IV drug abuse, and intraspinal abscess), which were added in previous years in response to prior reviews of the literature, feedback from the contractor's technical expert panel, and comments provided by other stakeholders.

1a.4 OTHER SOURCE OF EVIDENCE

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

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 0514
Corresponding Measures:
De.2. **Measure Title:** MRI Lumbar Spine for Low Back Pain
Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services
De.3. **Brief Description of Measure:** This measure evaluates the percentage of magnetic resonance imaging (MRI) of the lumbar spine studies for low back pain performed in the outpatient setting where conservative therapy was not attempted prior to the MRI. Antecedent conservative therapy may include claim(s) for physical therapy in the 60 days preceding the lumbar spine MRI, claim(s) for chiropractic evaluation and manipulative treatment in the 60 days preceding the lumbar spine MRI, or claim(s) for evaluation and management at least 28 days but no later than 60 days preceding the lumbar spine MRI. The measure is calculated based on a one-year window of Medicare claims data. The measure has been publicly reported, annually, by the measure steward, the Centers for Medicare & Medicaid Services (CMS), since 2010, as a component of its Hospital Outpatient Quality Reporting (HOQR) Program.
1b.1. **Developer Rationale:** This measure will reduce overuse of imaging for uncomplicated low back pain without prior attempts at antecedent conservative therapy, as overuse in this population can result in detection of incidental findings and reflect poor care coordination. The measure score will guide patient selection of providers, assess quality, and inform quality improvement.

S.4. **Numerator Statement:** MRI of the lumbar spine studies with a diagnosis of low back pain (from the denominator) without the patient having claims-based evidence of prior antecedent conservative therapy.
S.7. **Denominator Statement:** The number of MRI of the lumbar spine studies with a diagnosis of low back pain on the imaging claim performed in a hospital outpatient department on Medicare FFS beneficiaries within a 12-month time window.
S.10. **Denominator Exclusions:** Below, in Section S.11 we provide a detailed list of denominator exclusion conditions. Denominator exclusions are consistent with current guidelines, evidence in literature, and guidance from the measure TEP.

De.1. **Measure Type:** Process
S.23. **Data Source:** Claims (Only)
S.26. **Level of Analysis:** Facility, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Oct 28, 2008 **Most Recent Endorsement Date:** Oct 28, 2008

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not applicable; this is not a paired or grouped measure.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_0514_MeasureEvidenceForm.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure will reduce overuse of imaging for uncomplicated low back pain without prior attempts at antecedent conservative therapy, as overuse in this population can result in detection of incidental findings and reflect poor care coordination. The measure score will guide patient selection of providers, assess quality, and inform quality improvement.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Analysis of Medicare fee-for-service (FFS) claims data indicates variation in the use of inappropriate MRI lumbar spine studies. For the period from July 2014 to June 2015, performance rates ranged from 14.9 percent to 64.8 percent, with a weighted mean of 39.5 percent.

The data presented below represent information for the 1,128 facilities whose denominator counts met minimum case count requirements for all years included in the table.

Further details on the descriptive statistics for longitudinal facility performance are included below:

	2011*	2012*	2013*	2014**	2015**	2016**
Measurement Period	Jan 2009–Dec 2009			Jan 2010–Dec 2010	Jan 2011–Dec 2011	Jul 2012–Jun 2013
Facilities	1,128	1,128	1,128	1,128	1,128	1,128
Minimum Value	17.9%	12.3%	17.1%	17.6%	21.8%	14.9%
5th Percentile	23.4%	26.7%	27.0%	28.0%	30.4%	29.1%
25th Percentile	28.4%	32.0%	32.1%	33.1%	35.9%	35.3%
Median	31.9%	35.8%	35.9%	36.8%	39.8%	39.0%
75th Percentile	35.9%	40.1%	39.8%	41.0%	44.4%	43.5%
95th Percentile	43.5%	48.6%	48.5%	48.0%	51.8%	50.6%
Maximum Value	63.5%	69.1%	67.6%	67.7%	72.5%	64.8%
Mean Performance (Standard Deviation)	40.3% (6.6)	39.5% (6.6)		32.5% (6.2)	36.4% (6.8)	36.5% (6.6) 37.2% (6.2) Jul 2013–Jun 2014 Jul 2014–Jun 2015

*The measurement period for HOQR data reported from 2011 through 2013 ran from January to December. **Beginning with 2014 public reporting, the measurement period for HOQR was adjusted to run from July to June; consequently, data are not reported for January through June 2012.

The intentions for reporting this measure is to identify facilities with significant outlying performance and to reduce variation. As shown in the table above, significant outlying performance persists among wide variation, indicating there are facilities for which there is a notable rate of overuse.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Data have been included in Section 1b.2; these data represent national performance over time, from 2009 to 2015.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Using 2013 performance data, we evaluated the effect of patient and facility characteristics on the likelihood of each beneficiary having an inappropriate MRI lumbar spine study. Using a logistic regression model, we assessed the relationship between patient and facility characteristics for the 207,573 MRI lumbar spine studies performed in 2013 and found that beneficiary age, gender, and race, as well as facility characteristics, had a significant association with the rate of inappropriate MRI lumbar spine studies.

The regression model indicates that patient race/ethnicity is associated with inappropriate imaging. Asians were less likely to undergo inappropriate imaging compared to White beneficiaries (OR 0.866, p=0.020). There was no statistical differences for other racial groups compared to White beneficiaries.

The regression model also shows that gender is associated with inappropriate imaging; women were less likely to undergo inappropriate imaging compared to men (OR 0.848, p=0.000).

Patient age also had a statistically significant association with imaging use. When looking at Medicare FFS data, comparing to beneficiaries aged 60 to 69, beneficiaries aged 18 to 29 (OR 0.840, p=0.003), 40 to 49 (OR 0.898, p=0.000), 50 to 59 (OR 0.894, p=0.000), 70 to 79 (OR 0.874, p=0.000), and 80 to 89 (OR 0.827, p=0.000) were statistically less likely to receive an MRI lumbar spine study without appropriate antecedent conservative therapy. There was no statistical difference in the likelihood that a patient received an inappropriate MRI lumbar spine study for patients aged 30 to 39 or patients aged 90+ compared to beneficiaries aged 60 to 69.

Facility characteristics were also associated with rates of inappropriate imaging. When compared to facilities with fewer than 50 beds (a proxy for facility size), facilities with 101 to 250 beds were less likely to perform inappropriate MRI lumbar spine studies (OR 0.933, p=0.001). Similarly, a facility's urbanicity impacted a beneficiary's likelihood of having an inappropriate MRI lumbar spine study – urban facilities were less likely than rural facilities, likely caring for rural beneficiaries, to perform inappropriate MRI lumbar spine studies (OR 0.967, p=0.008). Finally, major-teaching facilities were less likely to perform inappropriate MRI lumbar spine studies (OR 0.890, p=0.000) compared to non-teaching facilities.

While the regression model identified subpopulations of patients and facilities for which there are statistically significant differences in the rate of inappropriate MRI lumbar spine studies, these disparities do not indicate a need for adjustment of the measure specifications. Adjusting for these differences would mask underlying differences in quality of care. As this is a process measure, there should be no difference in the standard of care for these patients; we believe these statistically significant differences are driven by variation in provider

practice. Consequently, we do not believe risk adjustment or stratification is necessary or appropriate for this measure.

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

The following literature excerpts highlight disparities associated with overuse of imaging for low back pain between genders, racial groups, and age bands.

Graves et al. (2012) assessed the association of demographic, job-related, psychosocial, and clinical factors with the use of MRI within six weeks from injury among workers' compensation claimants with acute low back pain for 1,830 workers. A total of 362 (19.8 percent) received an early MRI. Results of a multivariable regression demonstrated that male workers were 43 percent more likely to receive an early MRI than were female workers.

Mathias et al. (2012) used data from the HOQR Program to assess consistency in performance across measures, focusing on whether higher imaging use could be associated with certain hospital characteristics. To do so, the study team examined associations between hospital characteristics and higher use of imaging, drawing on 2008 HOQR data linked with data from the 2009 American Hospital Association survey. Mathias and his team found that use of imaging varied widely and was weakly correlated across most measures. Of note, hospitals with low volume (<25th percentile) were more likely to report higher imaging use than were hospitals of medium volume (25th to 75th percentile). Of particular interest, rural hospitals were more likely to report highest-decile use of lumbar spine MRI, in addition to several other measures. The study authors concluded that there are significant variations in use of imaging, with some hospitals reporting exceptionally high use.

Pham et al. (2009) analyzed Medicare claims from 2000 through 2002 and from 2004 through 2006, for 35,039 fee-for-service (FFS) Medicare beneficiaries with acute LBP. The research team found that minority beneficiaries received less rapid and less advanced imaging than did white beneficiaries. Beneficiaries covered by Medicaid also received less rapid and less advanced imaging when compared to other patients (22.7 percent versus 29.7 percent [p<.001] for imaging within 28 days, and 7.3 percent versus 11.0 percent [p<.001] for CT/MRI).

Friedman et al. (2010), using data from the National Hospital Ambulatory Medical Care Survey (NHAMCS), highlighted the frequency of emergency department (ED) visits for the treatment of low back pain and identified the diagnostic and therapeutic strategies used by physicians in a large sample representative of all ED visits throughout the United States. Results showed that, of all patients with LBP, 9.6 percent (95 percent CI: 7.2, 12.6) had a CT or MRI in 2006, compared with 3.2 percent (95 percent CI: 2.0, 5.1) in 2002 (p<0.01). Age and type of insurance were associated with advanced imaging, while geographic region was not.

Despite evidence identified in the literature indicating disparities in care for certain facility types and patient populations who present with LBP, these disparities do not indicate a need for adjustment of the measure specifications. Adjusting for these differences would mask underlying differences in quality of care. As this is a process measure that is not currently risk-adjusted, there should be no difference in the standard of care for these patients; we believe these statistically significant differences are driven by variation in provider performance. Consequently, we do not believe risk adjustment or stratification is necessary or appropriate for this measure.

REFERENCES

1.) Friedman BW, Chilstrom M, Bijur PE, Gallagher EJ. Diagnostic testing and treatment of low back pain in United States emergency departments: a national perspective. Spine. 2010; 35(24):E1406-11.
2.) Graves JM, Fulton-Kehoe, D, Martin DP, et al. Factors associated with early magnetic resonance imaging utilization for acute occupational low back pain: a population-based study from Washington State Workers´ Compensation. Spine. 2012; 37(19): 1708-1718.

3.) Mathias JS, Feinglass J, Baker DW. Variations in US hospital performance on imaging-use measures. Med Care. 2012;50(9):808-14.
4.) Pham HH, Landon BE, Reschovsky JD, et al. Rapidity and modality of imaging for acute low back pain in elderly patients. Archives of Internal Medicine. 2009; 169(10):972-981.

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

The measure addresses:

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

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Other
1c.2. **If Other:** Safety

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
There is consensus in the published literature that MRI for low back pain represents an area of significant inappropriate use. Acute LBP, with or without radiculopathy, is one of the most common health problems in the United States and is the leading cause of disability for individuals younger than 45; according to the American College of Radiology, uncomplicated acute low back pain is a benign, self-limited condition that warrants no imaging studies (Davis 2011). A thorough medical history and physical examination can usually identify the cause of low back pain. When low back pain is not attributed to potentially serious spinal pathology or non-spinal pathology, there is a poor correlation of imaging findings with low-back problems (Borenstein 2001; Carragee 2006; Jarvik 2003; Modic 2005; Maus 2010). Practitioners should emphasize that acute low back pain is nearly always benign and generally resolves itself within one to six weeks (Toward Optimized Practice 2011).

Despite consensus that there is little value in diagnostic imaging for acute LBP, the diagnosis is associated with significant practice variation in hospital outpatient settings for many imaging modalities, including X-ray imaging, discography, computed tomography (CT), MRI, electromyography, bone scans, thermography, and ultrasound imaging (Modic 2005; Chou 2007). Such use has enormous cost implications, largely due to the high cost of imaging studies and specialty referrals (Rao 2002). The cost of evaluating and treating acute low back pain runs into billions of dollars annually, not including time lost from work (Jenkins 2015). To evaluate the overuse of MRI in diagnosing LBP, Graves et al. (2014) estimated health care utilization and costs associated with adherence to clinical practice guidelines for the use of early MRI (within the first six weeks of injury) for acute occupational LBP. The study team identified workers (age > 18) with work-related low back pain using administrative claims and compared health care utilization and costs among workers whose imaging was adherent to guidelines (no early MRI) to workers whose imaging was not adherent to guidelines (early MRI in the absence of red flags). The study team found that, of 1,770 workers, the care of 336 (19.0 percent) was classified as non-adherent to guidelines. Outpatient and physical/occupational therapy utilization was 52 to 54 percent higher for workers whose imaging was not adherent to guidelines compared to workers with guideline-adherent imaging.

Webster et al. (2014) conducted a retrospective cohort study to compare type, timing, and longitudinal medical costs incurred after adherent versus non-adherent MRI for work-related LBP. The study examined a longitudinal workers´ compensation administrative data source to select low back pain claims filed between January 1, 2006 and December 31, 2006. Cases were grouped by MRI timing (early, timely, and no MRI) and sub-grouped by severity ("less severe" and "more severe"). The study team found that, for the cohort of 3,022, the adjusted

relative risks for MRI group cases to receive electromyography, nerve conduction testing, advanced imaging, injections, and surgery within six months post-MRI ranged from 6.5 (95% CI: 2.20-19.09) to 54.9 (95% CI: 22.12-136.21) times the rate for the referent group (no MRI, less severe). Medical costs for both early MRI subgroups were highest and increased the most over time. The investigators concluded that the impact of non-adherent MRI includes a wide variety of expensive and potentially unnecessary services, and occurs relatively soon post-MRI.

Webster and Cifuentes (2010) examined early MRI utilization for workers' compensation cases with acute, disabling low back pain as well as low or high propensity to undergo early MRI with disability duration, medical costs, and surgery. In a two-year follow-up study of 3,264 cases, the study team used Cox regression and generalized linear models to determine the association between both early MRI (first 30 days post-onset) and propensity of belonging to the early MRI group (estimated by demographic and severity indicators) with outcomes. The study team found that 21.7 percent cases had an early MRI. After controlling for covariates, cases that had early MRI and simultaneously had a low propensity to undergo early MRI were more likely to have worse outcomes. The investigators concluded that the majority of cases had no early MRI indications. Results suggest that the iatrogenic effects of early MRI are worse for disability and increase medical costs and surgery, unrelated to severity.

Weigel et al. (2012) evaluated the coordination of care for patients with low back pain whose services were supplemented by chiropractors. Using Medicare Part B claims from 1991 to 2007, the study team linked the claims data to chiropractic survey data obtained from interviews collected via the Assets and Health Dynamics among the Oldest Old survey. Results from Weigel's research showed substantial variation in the number and duration of chiropractic care, ranging from 3.74 to 23.12 episodes of care over 4.7 to 28.8 days. Over the 17-year study period, the study team found from 4.9 percent to 10.9 percent of patients had both chiropractic and non-chiropractic evaluation and management of their LBP. The investigators concluded that treatment for low back pain was frequently sought from a variety of provider types, but that there was little coordination of care across these providers in treating patients' pain.

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

1.) Borenstein DG, O'Mara JW, Boden SD, et al. The value of magnetic resonance imaging of the lumbar spine to predict low back pain in asymptomatic subjects: a seven-year follow-up study. American Journal of Bone and Joint Surgery. 2001; 83-A(9):1306-11.
2.) Carragee E, Alamin T, Cheng I, et al. Are first-time episodes of serious low back pain associated with new MRI findings? Spine. 2006; 6(6):624-35. Epub 2006 Oct 11.
3.) Chou R, Qaseem A, Snow V, Casey D, et al. Diagnosis and Treatment of Low Back Pain: A Joint Clinical Practice Guideline from the American College of Physicians and the American Pain Society. Ann Intern Med. 2007;147:478-491.
4.) Davis PC, Wippold FJ II, Cornelius RS, Angtuaco EJ, et al. Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria: low back pain. [online publication]. Reston (VA): American College of Radiology (ACR); 2011. 8 p. [48 references]
5.) Graves JM, Fulton-Kehoe D, Jarvik JG, et al. Health care utilization and costs associated with adherence to clinical practice guidelines for early magnetic resonance imaging among workers with acute occupational low back pain. Health Serv Res. 2014; 49(2): 645-665.
6.) Jarvik JG, Hollingworth W, Martin B, et al. Rapid magnetic resonance imaging vs. radiographs for patients with low back pain: a randomized controlled trial. Journal of the American Medical Association. 2003; 289(21):2810-8.
7.) Jenkins, H. J., et al. "Effectiveness of interventions designed to reduce the use of imaging for low back pain: a systematic review." Cmaj. 2015; 187(6): 401-408.
8.) Luo X, Pietrobon R, Sun SX, et al. Estimates and patterns of direct health care expenditures among individuals with back pain in the United States. Spine. 2004; 29:79–86.
9.) Maus, T. "Imaging the back pain patient." Phys Med Rehabil Clin N Am. 2010; 21(4): 725-766.
10.) Modic MT, Obuchowski NA, Ross JS, et al. Acute low back pain and radiculopathy: MR imaging findings and their prognostic role and effect on outcome. Radiology. 2005; 237(2):597-604.

11.) Rao JK, Kroenke K, Mihaliak KA, et al. Can guidelines impact the ordering of magnetic resonance imaging studies by primary care providers for low back pain? American Journal of Managed Care. 2002; 8(1):27-35.
12.) Toward Optimized Practice. Guideline for the evidence-informed primary care management of low back pain. Edmonton (AB): Toward Optimized Practice; 2011. 37 p. [39 references]
13.) Webster BS, Choi Y, Bauer AZ, et al. The cascade of medical services and associated longitudinal costs due to nonadherent magnetic resonance imaging for low back pain. Spine (Phila Pa 1976). 2014; 39(17): 1433-1440.
14.) Webster BS, Cifuentes M. Relationship of early magnetic resonance imaging for work-related acute low back pain with disability and medical utilization outcomes. J Occup Environ Med. 2010; 52(9): 900-7.
15.) Weigel PA , Hockenberry JM, Bentler SE, Kaskie B, et al. Chiropractic episodes and the co-occurrence of chiropractic and health services use among older Medicare beneficiaries. J Manipulative Physiol Ther. 2012; 35(3):168-175.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
This measure is not a PRO-PM measure.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

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

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Musculoskeletal, Musculoskeletal : Low Back Pain

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

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228695266120

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_0514_MeasureCodeList.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

As part of the annual measures maintenance and review process, several modifications were made to the list of excluded procedures and diagnoses, as well as to the look-back periods for certain diagnoses. In 2012, we added lumbar-spine surgery within the 90 days prior to the imaging procedure to the list of excluded procedures; prior lumbar-spine surgery is an appropriate reason for performing an MRI lumbar-spine study based on evidence in the literature (indicated in American College of Radiology Practice Guideline for the Performance of MRI of the Adult Spine). Look-back periods for each of the excluded clinical conditions were also added to the measure's specifications in 2012 (as identified by Lewin clinical staff and evaluated by the Technical Expert Panel [TEP]). Finally, nine additional exclusion categories (congenital spine and spinal-cord malformations, inflammatory and autoimmune disorders, infectious conditions, spinal vascular malformations and/or occult subarachnoid hemorrhage causes, spinal-cord infarction, effects of radiation, spinal abnormalities associated with scoliosis, syringohydromyelia, and postoperative fluid collections and soft-tissue changes) based on the 2012 update to the American College of Radiology Practice Guideline for the Performance of MRI of the Adult Spine and are aligned with the 2015 update to the ACR guideline. When presented to the contractor's TEP, the TEP supported these additions to the measure specifications.

There have been no changes in the measure specifications since the last measure update; however several measure refinements have been made in the past due to literature identified in CMS contractor's annual review of measure's evidence base and feedback from stakeholders.
Since the measure was initially endorsed in 2008, changes to the specifications include updates to the trauma exclusion and to the numerator evaluation and management (E&M) structure (in 2011); addition of the 90-day lookback period for the lumbar spine surgery exclusion (in 2012); and, addition of congenital spine/spinal cord malformations, inflammatory and autoimmune disorders, infectious conditions, spinal vascular malformations, spinal cord infarctions, effects from radiation, spinal abnormalities associated with scoliosis, syringohydromyelia, and postoperative fluid collections/soft tissue changes (in 2014).

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
MRI of the lumbar spine studies with a diagnosis of low back pain (from the denominator) without the patient having claims-based evidence of prior antecedent conservative therapy.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator: MRI lumbar spine studies with no evidence of antecedent conservative therapy (chiropractory or physical therapy within 60 days of the MRI study or an evaluation and management visit within 28 days to 60 days of the MRI study), for patients with LBP, performed within a 12-month time window.
Denominator: MRI lumbar spine studies, for patients with LBP, performed within a 12-month time window.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
For MRI lumbar-spine studies in the denominator, the numerator is defined by the following categories of antecedent conservative therapy:
-Claim(s) for physical therapy in the 60 days preceding the lumbar spine MRI
-Claim(s) for chiropractic evaluation and manipulative treatment in the 60 days preceding the lumbar spine MRI
-Claim(s) for evaluation and management =28 days and =60 days preceding the lumbar spine MRI

(Specific CPT codes for each type of antecedent conservative therapy are included in the value set for this measure; this detailed list can be found in the Excel workbook provided for criterion S2b.)

Time Period for Data: MRI lumbar spine studies with no evidence of antecedent conservative therapy (chiropractory or physical therapy within 60 days of the MRI study or an evaluation and management visit within 28 days to 60 days of the MRI study), for patients with low back pain, performed within a 12-month time window.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The number of MRI of the lumbar spine studies with a diagnosis of low back pain on the imaging claim performed in a hospital outpatient department on Medicare FFS beneficiaries within a 12-month time window.

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

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The denominator is defined by the following CPT codes:
MRI Lumbar Spine
CPT 72148, 72149, 72158

MRI lumbar spine CPT codes should be accompanied by a diagnosis of low back pain on the same claim:
ICD-9 codes 721.3, 721.90, 722.10, 722.52, 722.6, 722.93, 724.02, 724.2, 724.3, 724.5, 724.6, 724.70, 724.71, 724.79, 738.5, 739.3, 739.4, 846.1, 846.2, 846.3, 846.4, 846.8, 846.9, 847.2
ICD-10 codes M43.20, M43.25-M43.28, M43.5X5-M43.5X9, M43.8X5-M43.8X9, M43.9, M46.46-M46.47, M47.20, M47.26-M47.28, M47.816-M47.819, M47.896-M47.9, M48.06-M48.07, M51.26-M51.27, M51.34-M51.37, M51.86-M51.87, M53.2X7-M53.2X8, M53.3, M53.86-M53.88, M54.30-M54.32, M54.40-M54.42, M54.5, M54.89, M54.9, M99.03-M99.04, M99.23, M99.33, M99.43, M99.53, M99.63, M99.73, M99.83-M99.84, S33.5XXA-S33.9XXS
The diagnosis of low back pain must be on the MRI lumbar-spine claim (i.e., the lumbar-spine MRI must be billed with a low back pain diagnosis in one of the diagnoses fields on the claim). MRI lumbar spine studies without a diagnosis of low back pain on the claim are not included in the denominator count. If a patient had more than one MRI lumbar spine study for a diagnosis of low back pain on the same day only one study would be counted, but if a patient had multiple MRI lumbar spine studies with a diagnosis of low back pain on a claim during the measurement period each study would be counted (i.e., a patient can be included in the denominator count more than once).

Global and TC claims are considered in order to capture all outpatient volume facility claims, typically paid under the Outpatient Prospective Payment System (OPPS)/Ambulatory Payment Classifications (APC) methodology, and to avoid double counting of professional component claims (i.e., 26 modifier).

A technical unit can be identified by a modifier code of TC. A global unit can be identified by the absence of a TC or 26-modifier code.

MRI lumbar spine studies can be billed separately for the technical and professional components, or billed globally, which includes both the professional and technical components.

Professional component claims will outnumber TC claims due to over-reads.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

Below, in Section S.11 we provide a detailed list of denominator exclusion conditions. Denominator exclusions are consistent with current guidelines, evidence in literature, and guidance from the measure TEP.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Indications for measure exclusion include any patients with the following diagnosis code categories:
-Patients with lumbar spine surgery in the 90 days prior to MRI
-Cancer (within twelve months prior to MRI procedure)
-Congenital spine and spinal cord malformations (within five years prior to MRI procedure)
-Inflammatory and autoimmune disorders (within five years prior to MRI procedure)
-Infectious conditions (within one year prior to MRI procedure)
-Spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage (within five years prior to MRI procedure)
-Spinal cord infarction (within one year prior to MRI procedure)
-Neoplastic abnormalities (within five years prior to MRI procedure)
-Treatment fields for radiation therapy (within five years prior to MRI procedure)
-Spinal abnormalities associated with scoliosis (within five years prior to MRI procedure)
-Syringohydromyelia (within five years prior to MRI procedure)
-Postoperative fluid collections and soft tissue changes (within one year prior to MRI procedure)
-Trauma (within 45 days prior to MRI procedure)
-IV drug abuse (within twelve months prior to MRI procedure)
-Neurologic impairment: (within twelve months prior to MRI procedure)
-HIV (within twelve months prior to MRI procedure)
-Unspecified immune deficiencies (within twelve months prior to MRI procedure)
-Intraspinal abscess (an exclusion diagnosis must be in one of the diagnoses fields on the MRI lumbar spine claim)

(Specific CPT codes, ICD-9 codes, and ICD-10 codes for exclusion are included in the value sets for this measure; this detailed list can be found in the Excel workbook provided for criterion S2b.)

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not applicable; this measure does not stratify its results.

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

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

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

Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
No risk model specifications are provided, as risk adjustment or stratification is not necessary for this measure.

S.16. **Type of score:**
Other (specify):
If other: Percentage

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
This measure calculates the percentage of lumbar-spine MRI studies with a diagnosis of low back pain on the imaging claim for which the patient did not have prior claims-based evidence of antecedent conservative therapy. The measure is calculated based on hospital outpatient claims data, as follows:
1. Select hospital outpatient claims with a CPT code for any MRI lumbar-spine study on a revenue line item
2. Exclude professional component only claims with modifier = ´26´
3. Of claims identified in step 2, review relevant look-back periods for claims-based evidence of any procedure or diagnosis excluded from the measure; remove claims for which an exclusion has been identified
4. Set denominator counter = 1
5. Of claims identified in step 4, identify those claims for which there is no evidence of prior conservative therapy (claims for physical therapy in the 60 days preceding the imaging study; claims for chiropractic evaluation in the 60 days preceding the imaging study; or, claims for evaluation and management of at least 28 but equal to or less than 60 days prior to the imaging study). Set numerator count=1 for these claims
6. Aggregate denominator and numerator counts by facility identifier
7. Measure = numerator counts / denominator counts [The value should be recorded as a percentage]

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

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure relies exclusively on 100% Medicare FFS Standard Analytical File (SAF) data; no sampling of beneficiaries was performed.

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

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
The measure development 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.

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

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is not a PRO-PM measure.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : Regional and State

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

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

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_0514_MeasureTestingForm.docx

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.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

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

AND

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

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

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

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

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

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

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

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

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

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

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ ☒ administrative claims	☒ ☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: 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

a. 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.

b. 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.

c. 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 <u>performance measure score</u> 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*

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.

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
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.[2] 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).

[2] 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.

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[3] 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:

---

[3] 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%.

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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
This measure is claims-based, and uses CMS hospital outpatient claims as its data source.

Special attention needs to be taken when counting procedures on the Medicare claims files. The biggest issue is how to deal with modifier codes. Modifiers are two digit indicators (alpha or numeric) that represent a service or procedure that has been altered by some specific circumstance, which typically will impact the payment amount.

Procedure modifier code "26" represents the professional component of a procedure and includes the clinician work (i.e., the reading of the image by a physician), associated overhead and professional liability insurance costs. This modifier corresponds to the human involvement in a given service or procedure.

The procedure modifier code "TC" represents the technical component of a service or procedure and includes the cost of equipment and supplies to perform that service or procedure. This modifier corresponds to the equipment/facility part of a given service or procedure.

In most cases, unmodified codes represent a global procedure which includes both the professional and technical components. There are also other modifier codes. All other modifier codes have been counted as a technical code for our purposes. When calculating the measures, we are only concerned with procedures associated with technical and global modifiers, as these modifiers refer to services provided by the facility. This reduces the possibility of double-counting procedures, since a single procedure may result in both a technical and professional record on the claims files. There were very few instances when this occurred as it related to procedures applicable to the measure.

When developing counts of procedures, the objective is to avoid double-counting procedures that may have been billed through multiple revenue centers within a facility. Billing through multiple centers leads to multiple records in the Medicare claims files (i.e., the SAFs). For instance, there may be multiple bills for a single MRI. On one bill, the charges relate to the application of a radiopharmaceutical, which could have a technical modifier code and come from the pharmacy revenue center. On the other bill, the charges relate to the imaging study and may fall under a technical bill from the imaging center revenue center. In this case, we only count the MRI once, since only one MRI was performed. However, if we were summing up the Medicare paid amounts for this procedure, we would include the Medicare paid amounts from both bills, as they each represent payments for services directly related to the particular MRI procedure.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
No fees, licensure, or other requirements are necessary to use this measure; however, CPT codes, descriptions, and other data are copyright 2015 American Medical Association. All rights reserved. CPT is a registered trademark of the American Medical Association. Applicable FARS\DFARS Restrictions Apply to Government Use. Fee schedules, relative value units, conversion factors, and/or related components are not assigned by the AMA, are not part of CPT, and the AMA is not recommending their use. The AMA does not directly or indirectly practice medicine or dispense medical services. The AMA assumes no liability for data contained or not contained herein.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Hospital Outpatient Quality Reporting http://www.medicare.gov/hospitalcompare/search.html Hospital Outpatient Quality Reporting https://www.qualitynet.org/dcs/ContentServer? c=Page&pagename=QnetPublic% 2FPage% 2FQnetTier2&cid=1228695266120 Hospital Outpatient Quality Reporting http://www.medicare.gov/hospitalcompare/search.html Hospital Outpatient Quality Reporting https://www.qualitynet.org/dcs/ContentServer? c=Page&pagename=QnetPublic% 2FPage% 2FQnetTier2&cid=1228695266120 Quality Improvement (external benchmarking to organizations) Hospital Outpatient Quality Reporting http://www.medicare.gov/hospitalcompare/search.html Hospital Outpatient Quality Reporting https://www.qualitynet.org/dcs/ContentServer? c=Page&pagename=QnetPublic% 2FPage% 2FQnetTier2&cid=1228695266120

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

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

Public Reporting:
Name of program and sponsor: The CMS Hospital Outpatient Quality Reporting (HOQR) Program
Purpose: The HOQR Program is a pay for quality data reporting program implemented by CMS for outpatient hospital services. In addition to providing hospitals with a financial incentive to report their quality of care measure data, the HOQR Program provides CMS with data to help Medicare beneficiaries make more informed decisions about their health care. Hospital quality of care information gathered through the HOQR Program is publicly available on the Hospital Compare website.
Accountable entities and patients: The publicly reported values (on Hospital Compare) are calculated for all facilities in the United States that meet minimum case count and other reporting requirements. For the period of 2009 to 2014, 1,189 facilities met the minimum case count each year. Additional facilities met the minimum case count requirements in some, but not all, years. The claims included in the publicly reported calculations are for Medicare FFS patients whose claims are subject to OPPS.
Level of measurement and setting: HOQR measures are measured at the facility, state, and national level. The setting for HOQR measures is the hospital outpatient care setting.

Quality Improvement with Benchmarking (external benchmarking to multiple organizations):
Name of program and sponsor: The CMS HOQR Program
Purpose: The HOQR Program is a pay for quality data reporting program implemented by CMS for outpatient hospital services. In addition to providing hospitals with a financial incentive to report their quality of care measure data, the data is publicly reported on the Hospital Compare website. The data reported on Hospital Compare not only shows the hospital's score on the measure, but also provides state and national averages for

the measure. This enables consumers to compare the hospital's performance to other facilities and determine if the facility is an outlier.
Accountable entities and patients: The publicly reported values (on Hospital Compare) are calculated for all facilities in the United States that meet minimum case count and other reporting requirements. For the period of 2009 to 2014, 1,189 facilities met the minimum case count each year. Additional facilities met the minimum case count requirements in some, but not all, years. The claims included in the publicly reported calculations are for Medicare FFS patients whose claims are subject to the OPPS.
Level of measurement and setting: HOQR measures are measured at the facility, state, and national level. The setting for HOQR measures is the hospital outpatient care setting.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This measure is publicly reported.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
This measure is publicly reported.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Summary statistics for performance scores from 2009 to 2015 are provided in Section 1b.2 for the hospital outpatient setting. The median rate of overuse for facilities increased from 2011 to 2015 (31.9% to 39.0%); however, performance has improved for subsets of facilities over the period of public reporting. For facilities that were classified as outliers in the first year of reporting (facilities with the highest 10% of performance scores, indicating poor performance) for the measure, only 16% were still outliers five years later. Facilities with persistent poor performance tend to be rural, small, and non-teaching.

Though national performance has not improved in the hospital outpatient setting since the inception of public reporting, an improvement in performance over the last two years suggests that facilities are beginning to address this gap in quality of care.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Though national performance has not improved in the hospital outpatient setting since the inception of public reporting, an improvement in performance over the last two years suggests that facilities are beginning to address this gap in quality of care.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
CMS conducts prepayment claims analysis and post-payment audits that should prevent this factor from having a major impact on the measure calculations performed on claims data.

5. Comparison to Related or Competing Measures

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

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

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0052 : Use of Imaging Studies for Low Back Pain

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
National Committee for Quality Assurance (NCQA) - Back Pain: Appropriate Imaging for Acute Back Pain
Institute for Clinical Systems Improvement (ICSI) - Adult acute and subacute low back pain: Percentage of patients with a diagnosis of non-specific back pain for whom the physician ordered imaging studies during the six weeks after pain onset, in the absence of red flags
ICSI - Adult acute and subacute low back pain: Percentage of patients with radicular pain for whom the clinician ordered imaging studies during the six weeks after pain onset

5a. Harmonization
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
NQF #0514 is similar in construct to NQF measure #0052, Use of Imaging Studies for Low Back Pain (developed by NCQA). Both measures consider the overuse of imaging for patients with a diagnosis of low back pain. However, the measures have key differences in intent and patient population that limit the desirability of

complete harmonization. NQF #0052 is a resource utilization measure and does not consider the administration of antecedent conservative therapy prior to imaging (captured in NQF #0514); NQF #0052 includes multiple imaging modalities (i.e., CT, MRI, and X-ray) and several regions of the spine (lumbar, cervical, and thoracic), while NQF #0514 focuses solely on MRI lumbar spine studies; and, NQF #0514 is calculated using Medicare claims data while NQF #0052 is calculated using both administrative claims and electronic clinical data. The stewards for NQF #0514 and #0052 have held a series of harmonization discussions, focusing on the clinical intents of the measures and looking for opportunities for alignment. To better harmonize the measures, the stewards have updated the value sets used to identify low back pain and exclusion conditions. The stewards recommend against complete harmonization due to differences in the clinical focus (only lumbar spine imaging vs. lumbar, cervical, and thoracic spine imaging), intent (inappropriate use vs. resource utilization), age of target population, and structure of the measures. While NQF #0514 is related to the second NCQA measure, significant structural differences make harmonization undesirable: the NCQA measure is calculated for individual physicians; the NCQA measure does not consider the administration of antecedent conservative therapy prior to imaging; the measure includes several imaging modalities, and; it focuses on a different patient population (patients aged 18-80) than NQF #0514. Similarly, while NQF #0514 is related to the ICSI measures, the measures have key differences that limit the desirability of complete harmonization. The first ICSI measure differs from NQF #0514 in several key ways: the measures target different patient populations, and; the ICSI measure does not consider antecedent conservative therapy prior to imaging. There are significant differences between the second ICSI measure and NQF #0514, including: targeted patient populations, and consideration (or not) of antecedent conservative therapy prior to imaging.

5b. Competing Measures

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

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

Appendix

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

Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services
Co.2 **Point of Contact:** Vinitha, Meyyur, Vinitha.Meyyur@cms.hhs.gov, 410-786-7224-
Co.3 **Measure Developer if different from Measure Steward:** The Lewin Group

Co.4 **Point of Contact:** Colleen, McKiernan, Colleen.McKiernan@lewin.com, 703-269-5595-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

The contractor has convened a TEP, which will evaluate and provide feedback on measure-development and maintenance efforts for the imaging efficiency measures. Specifically, the TEP provides direction and feedback through all phases of project activities including updates to the current specifications of the six imaging efficiency measures, review of quantitative testing results, feedback on qualitative testing questions (i.e., results of TEP member questionnaires), and support for endorsement of the measures by NQF.

The following is a list of the contractor's TEP members:

Meenu Arora, MBA
Quality Improvement Leader, Sequoia Hospital

Brian Baker
Chief Executive Officer, Carealytics

Peter Benner
Vice Chair, MNSure

Martha Deed, Ph.D
Safe Patient Project´s Patient Advocacy Network

Lawrence Feinberg, MD
Attending Physician, University of Colorado Hospital

Elliott Fishman, MD
Professor of Radiology and Oncology, Johns Hopkins School of Medicine

Marian Hollingsworth
Patient Advocate

Michael Hutchinson, MD, Ph.D
Clinical Associate Professor of Neurology, Icahn School of Medicine at Mount Sinai

Gregory M. Kusiak, MBA, FRBMA
President, California Medical Business Services, Inc.

Barbara Landreth, RN, MBA
Clinical Information Analyst , St. Louis Area Business Health Coalition

Barbara McNeil, MD, Ph.D
Head Professor of Radiology, Harvard University

Michael J. Pentecost, MD
Chief Medical Officer, NIA Magellan

David Seidenwurm, MD

Medical Staff Consultant, Sutter Medical Group

Adam Sharp, MD, MS
Research Scientist , Kaiser Permanente Southern California

Paul R. Sierzenski, MD, RDMS, FACEP, FAAEM
Medical Director, Christian Health Care System

C. Todd Staub, MD, FACP
Chairman, ProHealth Physicians

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2011

Ad.3 **Month and Year of most recent revision:** 03, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

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

Appendix B: Measure Evaluation Summaries

0052 Use of Imaging Studies for Low Back Pain

Submission

Description: The percentage of patients with a primary diagnosis of low back pain who did not have an imaging study (plain X-ray, MRI, CT scan) within 28 days of diagnosis.

Numerator Statement: Patients who received an imaging study (x-ray, CT, MRI) within the 28 days following a diagnosis of low back pain.

Denominator Statement: All patients 18 years as of January 1 of the measurement year to 50 years as of December 31 of the measurement year with a claim/encounter for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, with a principal diagnosis of low back pain during the Intake Period (January 1 – December 3 of the measurement year).

Exclusions: Because the intent of the measure is to assess imaging for patients with a new episode of low back pain, exclude patients with a recent diagnosis of low back pain.

Also, exclude any patient who had a diagnosis for which imaging is clinically appropriate. Any of the following meet criteria:

(1) Cancer
(2) Trauma
(3) Recent IV drug abuse
(4) Neurologic impairment
(5) HIV
(6) Spinal infection
(7) Major organ transplant
(8) Prolonged use of corticosteroids

Adjustment/Stratification: No risk adjustment or risk stratification

Level of Analysis: Health Plan; Integrated Delivery System

Setting of Care: Clinician Office/Clinic; Emergency Department; Urgent Care - Ambulatory

Type of Measure: Process

Data Source: Claims (Only); This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

Measure Steward: National Committee for Quality Assurance

STEERING COMMITTEE MEETING [12/12/16]

1. Importance to Measure and Report: The measure meets the Importance criteria

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

1a. Evidence: **Previous Evidence Evaluation Accepted**; 1b. Performance Gap: **H-0**; **M-16**; **L-3**; **I-0**

Rationale:

- The developer updated the evidence to include the 2015 American College of Radiology (ACR) Appropriateness Criteria: Low Back Pain. The Committee agreed that this was an appropriate update to the evidence and there was no need to re-discuss and re-vote on the evidence sub-criterion.
- Although the developers were unable to provide current performance rates for the measure as specified, the Committee agreed that rates of inappropriate imaging of patients with low back pain continues to be relatively high.

0052 Use of Imaging Studies for Low Back Pain

- The developer referenced a recent study from the Department of Veterans Affairs, which found significantly higher rates of MRI in younger adults compared to older adults and significantly lower rates in blacks compared to whites.

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

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

2a. Reliability: **H-0**; **M-1**; **L-0**; **I-17** 2b. Validity: **H-0**; **M-0**; **L-5**; **I-12**

<u>Rationale</u>:

- As a response to Committee feedback during the previous evaluation of the measure, the developer revised the specifications to expand the population being measured (i.e., including physical therapy and telehealth visits, shortening the look-back period for the exclusion due to recent trauma, and expanding the list of exclusions to include those with prolonged use of corticosteroids, HIV, major organ transplant, and recent spinal infection).
- Committee members had several questions and concerns about the measure specifications, as follows:
  - Members questioned not requiring pharmacy data to identify those with prolonged use of corticosteroids. The developers acknowledged that use of pharmacy claims could potentially identify additional patients who should be excluded from the measure, but noted that a requirement for the pharmacy benefit would reduce the number of plans that could report on the measure. They stated that, according to their testing information, the rate of prolonged corticosteroid use is quite low, suggesting that lack of pharmacy data does not substantively impact the measure results.
  - Members noted a preference for excluding any patient with a history of spine surgery. The developers noted that the measure can capture anyone with spine surgery in the 6 months prior to the measure index date.
  - Members asked whether patients with radicular symptoms would or would not be excluded from the measure. The developers noted and the Committee agreed that guidelines indicate these patients should not have imaging in the first six weeks. They clarified that these patients would not be excluded from the measure via the neurologic impairment exclusion (i.e., they would be included in the measure).
  - One member noted that spinal infection often is identified via imaging. The developers noted that if a patient has a diagnosis within 28 days of the imaging study that indicates spinal infection, that patient is excluded from the measure.
  - Committee members questioned the 28-day post-imaging threshold for exclusions, noting that the timeframe seems somewhat arbitrary and suggesting that it might not be long enough if LBP is treated with conservative management at initial diagnosis. The developers clarified that the 28-day threshold is applied after the imaging study, not after the initial diagnosis.
  - Committee members asked for clarifications about the timeframe for trauma exclusions and about what, specifically, is identified as "trauma." The developers noted that the value set included as part of the measure specifications include a list of codes used for the trauma exclusions. They also clarified that the timeframe for trauma exclusions is within three months prior to the diagnosis of LBP.
  - One member noted that the literature indicates that patients with known anatomic spinal anomalies may not require imaging, although this is not covered in the guidelines. The developers noted they have crafted the measure to align with current guidelines.
- Updated score-level testing results based on the revised specifications were not provided. The developer noted that these data would not be available for analysis until mid-2017.
- Although the developers did not provide updated data-element level validity testing, they did provide additional information from their 2002 field testing analysis; these data provided insight on the ability to identify patients in the recently-added exclusion categories (i.e., prolonged steroid use, spinal

0052 Use of Imaging Studies for Low Back Pain

infection, and immunosuppression). The developer stated that according to the administrative data from 2002, the various red-flag conditions specified as exclusions in the measure occur in 0% to 1.9% of LBP episodes (4.9% overall).

- Committee members expressed concern that a significant number of patients with trauma or neurologic impairment are not being captured using administrative claims data. The developer responded that the testing was performed using 2003-2004 data. They suggested the possibility that claims for trauma and neurologic impairment may have improved since then. They also noted a lack of feedback from health plans that the measure is actually missing a lot of trauma cases.
- In their submission materials, the developer presented their expert panel and commenting processes as an assessment of face validity. As part of the discussion, they provided additional explanation of how they believe these processes fulfill NQF's requirement for face validity.
- The Committee's rating of reliability had to depend, to a large extent, one the data element validity testing information provided by the developer (because the developer was unable to update their score-level reliability testing at this time). Therefore, NQF staff allowed for a vote on validity prior to the vote on reliability. Ultimately, the Committee agreed that there was insufficient information provided for validity and did not recommend the measure for endorsement.

STAFF NOTE: In the staff preliminary analysis and during the measure evaluation webinar, NQF staff noted that only percentage agreement statistics were provided to show the level of agreement between administrative codes and medical records and that the results provided were not calculated using the newly-specified measure. However, after further reflection on the additional information that was provided as part of the exclusion analysis (specifically, data on the ability to identify the new exclusions in claims only, in medical records only, or in both), staff determined that these data shed light on questions that are addressed in sensitivity/specificity analysis, even though the developer did not provide actual sensitivity/specificity statistics. Thus, staff no longer considers the data element validity testing presented by the developer to be insufficient. Because we have reversed our previous guidance, NQF will ask the Committee to reconsider and re-vote on Reliability and Validity during the post-comment call. Also, the developer is seeking updated data to present during the post-comment call.

3. Feasibility: H-12; M-7; L-0; I-0

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

STAFF NOTE: Even though this measure did not pass the Validity subcritieria, staff asked the Committee to use the remaining time on the webinar to discuss and vote on the Feasibility and Usability and Use criteria, given that the developer would be providing additional data during the post-comment call.

Rationale:

- All data elements are in defined fields in electronic claims and generated or collected by and used by healthcare personnel during the provision of care. The Committee agreed that the data are readily available and can be captured without undue burden.

4. Usability and Use: H-0; M-18; L-0; I-0

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

Rationale:

- This measure is being used in at least 10 accountability programs, including pay-for-performance programs, accreditation programs, and public reporting.
- One member asked if there might be unintended consequences of the measure if patients who meet red-flag conditions from the guideline are not being excluded from the measure. The developers acknowledged the possibility, but emphasized their process of obtaining feedback from plans.

0052 Use of Imaging Studies for Low Back Pain
• The Committee expressed concern about the lack of improvement in performance in the last several years. The developer expressed hope that the Choosing Wisely Campaign would help promote more attention to the issue and drive improvement in performance of the measure.

5. Related and Competing Measures
This measure is competing with:
• #0514: MRI Lumbar Spine for Low Back Pain (CMS)
  ○ Due to differences in the level of analysis and care settings, the Committee will not be asked to select a best in-class measure.
  ○ Since the last evaluation, the developers have worked to harmonize the measures, resulting in greater congruence in how low back pain is defined, how cancer exclusions are defined, and in exclusion categories. Areas where the measures are not yet harmonized include the specific imaging modalities captured by the measure and some of the exclusion categories.

Steering Committee Recommendation for Endorsement: **Not Recommended** <u>Rationale:</u> •
6. Public and Member Comment •
7. Consensus Standards Approval Committee (CSAC) Vote: Y-X; N-X
8. Appeals

0514 MRI Lumbar Spine for Low Back Pain

Submission

Description: This measure evaluates the percentage of magnetic resonance imaging (MRI) of the lumbar spine studies for low back pain performed in the outpatient setting where conservative therapy was not attempted prior to the MRI. Antecedent conservative therapy may include claim(s) for physical therapy in the 60 days preceding the lumbar spine MRI, claim(s) for chiropractic evaluation and manipulative treatment in the 60 days preceding the lumbar spine MRI, or claim(s) for evaluation and management at least 28 days but no later than 60 days preceding the lumbar spine MRI. The measure is calculated based on a one-year window of Medicare claims data. The measure has been publicly reported, annually, by the measure steward, the Centers for Medicare & Medicaid Services (CMS), since 2010, as a component of its Hospital Outpatient Quality Reporting (HOQR) Program.

Numerator Statement: MRI of the lumbar spine studies with a diagnosis of low back pain (from the denominator) without the patient having claims-based evidence of prior antecedent conservative therapy.

Denominator Statement: The number of MRI of the lumbar spine studies with a diagnosis of low back pain on the imaging claim performed in a hospital outpatient department on Medicare FFS beneficiaries within a 12-month time window.

Exclusions: Indications for measure exclusion include any patients with the following diagnosis code categories:

-Patients with lumbar spine surgery in the 90 days prior to MRI

-Cancer (within twelve months prior to MRI procedure)

-Congenital spine and spinal cord malformations (within five years prior to MRI procedure)

-Inflammatory and autoimmune disorders (within five years prior to MRI procedure)

-Infectious conditions (within one year prior to MRI procedure)

-Spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage (within five years prior to MRI procedure)

-Spinal cord infarction (within one year prior to MRI procedure)

-Neoplastic abnormalities (within five years prior to MRI procedure)

-Treatment fields for radiation therapy (within five years prior to MRI procedure)

-Spinal abnormalities associated with scoliosis (within five years prior to MRI procedure)

-Syringohydromyelia (within five years prior to MRI procedure)

-Postoperative fluid collections and soft tissue changes (within one year prior to MRI procedure)

-Trauma (within 45 days prior to MRI procedure)

-IV drug abuse (within twelve months prior to MRI procedure)

-Neurologic impairment: (within twelve months prior to MRI procedure)

-HIV (within twelve months prior to MRI procedure)

-Unspecified immune deficiencies (within twelve months prior to MRI procedure)

-Intraspinal abscess (an exclusion diagnosis must be in one of the diagnoses fields on the MRI lumbar spine claim)

(Specific CPT codes, ICD-9 codes, and ICD-10 codes for exclusion are included in the value sets for this measure; this detailed list can be found in the Excel workbook provided for criterion S2b.)

Adjustment/Stratification: No risk adjustment or risk stratification

Level of Analysis: Facility; Population : Regional and State

Setting of Care: Emergency Department, Clinician Office/Clinic, Hospital : Hospital, Hospital : Acute Care Facility, Hospital : Critical Care, Imaging Facility, Urgent Care - Ambulatory

Type of Measure: Process

Data Source: Claims (Only)

Measure Steward: Centers for Medicare and Medicaid Services

0514 MRI Lumbar Spine for Low Back Pain

STEERING COMMITTEE MEETING [1/6/2017]

1. Importance to Measure and Report: The measure meets the Importance criteria

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

1a. Evidence: **Previous Evidence Evaluation Accepted**; 1b. Performance Gap: **H-3**; **M-10**; **L-0**; **I-0**

Rationale:

- The developer updated the evidence to include the 2015 American College of Radiology (ACR) Appropriateness Criteria: Low Back Pain. The Committee agreed that this was an appropriate update to the evidence and there was no need to re-discuss and re-vote on the evidence sub-criterion.
- To demonstrate opportunity for improvement, the developer provided an analysis of Medicare fee-for-service (FFS) claims data that indicates variation in the use of inappropriate MRI lumbar spine studies. Performance rates for July 2104 to June 2015 averaged 39.5% and ranged from 14.9% to 64.8% (NOTE: a lower rate is better).
- Committee members noted that the performance gap data actually demonstrated a decrease in performance (from 32.5% in 2009 to 39.5% in 2014-2015). The developer indicated that this could be a result of a change in data sources that were used to compute performance scores. The developer also noted that changes in specifications over time make it difficult to interpret changes in performance across time (specifically, expanding the exclusions would decrease the measure denominator, but would not uniformly affect the measure result).
- 2013 data presented by the developer showed that beneficiary age, gender, and race, as well as facility characteristics (i.e., number of beds, urban/rural locality, teaching status) were significantly associated with the rate of inappropriate MRI lumbar spine studies.

2. Scientific Acceptability of Measure Properties: The measure does not meet the Scientific Acceptability criteria

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

2a. Reliability: **H-0**; **M-8**; **L-5**; **I-0** 2b. Validity: **H-0**; **M-3**; **L-9**; **I-1**

Rationale:

- Committee members had several questions and concerns about the measure specifications, as follows:
    - The measure is specified for Medicare Fee-for-Service beneficiaries. However, "elderly individuals" is one of the red-flag conditions in the Appropriate Use guideline, indicating that imaging for the patients presenting with LBP may be appropriate. The developer interpreted the guideline as indicating that "elderly" should not be an independent indicator for imaging; however, some Committee members disagreed with this interpretation.
    - The measure uses evaluation and management (E&M) visits as a proxy for antecedent conservative care (in addition to claims for physical therapy or chiropractic visits). In general, the Committee agreed that the E&M visits are a reasonable proxy for some kinds of antecedent therapy, but questioned whether they would capture other types of antecedent therapy such as telephone encounters. Members noted that some types of antecedent conservative care (e.g., NSAIDs, Tylenol, massage therapy, acupuncture) cannot be captured in claims data.
    - Members questioned several of the look-back periods for some of the exclusions (e.g., 90 days for spine surgery, 12 months for cancer; 5 years for congenital spine and spinal cord malformations). For congenital malformations, the developer clarified that the 5-year look-back was mainly because of lack of access to historical data.
    - Committee members expressed concern that specific codes for neurological impairment, specifically those for which the evidence supports appropriate use of MRI, are not adequately captured in this measure. The developer agreed to look into the coding, but also noted that the red flag conditions often occur in tandem, meaning individual patients often are excluded from the measure due to several of the existing measure exclusions. Committee members noted that sciatica radiculopathy, typically does not present with other red-flag conditions.

0514 MRI Lumbar Spine for Low Back Pain

- The Committee expressed confusion about what changes, if any, have been made to the measure since the 2014 evaluation. Although the developers described the various analysis they performed (e.g., quantitative and qualitative evaluation of the look-back periods for several of the measure exclusions), it was still not clear to the Committee if or how the measure has been revised. Some of the confusion dates back to the 2014 evaluation, when the developer had actually added several exclusions to the measure that were not apparent in the submission materials considered by the Committee.
- The developers presented updated score-level signal-to-noise reliability testing using 2013 data. Reliability scores from this analysis ranged from 22.4% to 86.6%, with a median reliability score of 44.9%. The median value was well below 0.7, which is often used as a rule-of-thumb minimal acceptable value, and lower than the 53.1% found in previous testing. The developers also provided, a couple of days prior to the evaluation webinar, another set of testing results. This new testing used a split-sample (or "test-retest") approach to compare agreement in performance across hospitals. The intraclass correlation coefficient from this analysis was 0.59, which can be interpreted as moderate agreement (i.e., there is moderate consistency in performance within facilities).
- The developers assessed the face validity of the measure score by surveying an 11-member Technical Expert Panel (TEP). They asked the TEP members to indicate whether the measure captures the most appropriate and prevalent types of antecedent conservative therapy available through claims data (8 of 11 said yes) and to indicate their agreement as to whether the measure helps assess the inappropriate use of MRI lumbar-spine tests (9 of 11 agreed or strongly agreed).
- The developer clarified that the intent of the measure is not to drive measure results to zero, but to decrease the number of orders for MRI on presentation of LBP and to reduce variation between facilities in inappropriate MRIs.
- After much discussion, the Committee agreed that the measure did not pass the validity subcriterion and did not recommend the measure for endorsement.

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

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

Rationale:

-

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

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

Rationale:

-

5. Related and Competing Measures

This measure is competing with:

- #0052: Use of Imagine Studies for Low Back Pain (NCQA)
  - Due to differences in the level of analysis and care settings, the Committee will not be asked to select a best in-class measure.
  - Since the last evaluation, the developers have worked to harmonize the measures, resulting in greater congruence in how low back pain is defined, how cancer exclusions are defined, and in exclusion categories. Areas where the measures are not yet harmonized include the specific imaging modalities captured by the measure and some of the exclusion categories.

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

Rationale:

-

6. Public and Member Comment

-

0514 MRI Lumbar Spine for Low Back Pain
7. Consensus Standards Approval Committee (CSAC) Vote: Y-X; N-X
8. Appeals

Appendix B: Details of Measure Evaluation

Rating Scale: H=High; **M**=Moderate; **L**=Low; **I**=Insufficient; **NA**=Not Applicable; **Y**=Yes; **N**=No

Measures Not Recommended

0052 Use of Imaging Studies for Low Back Pain

Submission

Description: The percentage of patients with a primary diagnosis of low back pain who did not have an imaging study (plain X-ray, MRI, CT scan) within 28 days of diagnosis.

Numerator Statement: Patients who received an imaging study (x-ray, CT, MRI) within the 28 days following a diagnosis of low back pain.

Denominator Statement: All patients 18 years as of January 1 of the measurement year to 50 years as of December 31 of the measurement year with a claim/encounter for an outpatient, observation, emergency department, physical therapy, or telehealth visit, or osteopathic or chiropractic manipulative treatment, with a principal diagnosis of low back pain during the Intake Period (January 1 – December 3 of the measurement year).

Exclusions: Because the intent of the measure is to assess imaging for patients with a new episode of low back pain, exclude patients with a recent diagnosis of low back pain.

Also, exclude any patient who had a diagnosis for which imaging is clinically appropriate. Any of the following meet criteria:

(1) Cancer
(2) Trauma
(3) Recent IV drug abuse
(4) Neurologic impairment
(5) HIV
(6) Spinal infection
(7) Major organ transplant
(8) Prolonged use of corticosteroids

Adjustment/Stratification: No risk adjustment or risk stratification

Level of Analysis: Health Plan; Integrated Delivery System

Setting of Care: Clinician Office/Clinic; Emergency Department; Urgent Care - Ambulatory

Type of Measure: Process

Data Source: Claims (Only); This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

Measure Steward: National Committee for Quality Assurance

STEERING COMMITTEE MEETING [12/12/16]

1. Importance to Measure and Report: The measure meets the Importance criteria

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

1a. Evidence: **Previous Evidence Evaluation Accepted**; 1b. Performance Gap: **H-0**; **M-16**; **L-3**; **I-0**

Rationale:

- The developer updated the evidence to include the 2015 American College of Radiology (ACR) Appropriateness Criteria: Low Back Pain. The Committee agreed that this was an appropriate update to

the evidence and there was no need to re-discuss and re-vote on the evidence sub-criterion.

- Although the developers were unable to provide current performance rates for the measure as specified, the Committee agreed that rates of inappropriate imaging of patients with low back pain continues to be relatively high.
- The developer referenced a recent study from the Department of Veterans Affairs, which found significantly higher rates of MRI in younger adults compared to older adults and significantly lower rates in blacks compared to whites.

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

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

2a. Reliability: **H-0**; **M-1**; **L-0**; **I-17** 2b. Validity: **H-0**; **M-0**; **L-5**; **I-12**

<u>Rationale</u>:

- As a response to Committee feedback during the previous evaluation of the measure, the developer revised the specifications to expand the population being measured (i.e., including physical therapy and telehealth visits, shortening the look-back period for the exclusion due to recent trauma, and expanding the list of exclusions to include those with prolonged use of corticosteroids, HIV, major organ transplant, and recent spinal infection).
- Committee members had several questions and concerns about the measure specifications, as follows:
  - Members questioned not requiring pharmacy data to identify those with prolonged use of corticosteroids. The developers acknowledged that use of pharmacy claims could potentially identify additional patients who should be excluded from the measure, but noted that a requirement for the pharmacy benefit would reduce the number of plans that could report on the measure. They stated that, according to their testing information, the rate of prolonged corticosteroid use is quite low, suggesting that lack of pharmacy data does not substantively impact the measure results.
  - Members noted a preference for excluding any patient with a history of spine surgery. The developers noted that the measure can capture anyone with spine surgery in the 6 months prior to the measure index date.
  - Members asked whether patients with radicular symptoms would or would not be excluded from the measure. The developers noted and the Committee agreed that guidelines indicate these patients should not have imaging in the first six weeks. They clarified that these patients would *not* be excluded from the measure via the neurologic impairment exclusion (i.e., they would be included in the measure).
  - One member noted that spinal infection often is identified via imaging. The developers noted that if a patient has a diagnosis within 28 days of the imaging study that indicates spinal infection, that patient is excluded from the measure.
  - Committee members questioned the 28-day post-imaging threshold for exclusions, noting that the timeframe seems somewhat arbitrary and suggesting that it might not be long enough if LBP is treated with conservative management at initial diagnosis. The developers clarified that the 28-day threshold is applied after the imaging study, not after the initial diagnosis.
  - Committee members asked for clarifications about the timeframe for trauma exclusions and about what, specifically, is identified as "trauma." The developers noted that the value set included as part of the measure specifications include a list of codes used for the trauma exclusions. They also clarified that the timeframe for trauma exclusions is within three months prior to the diagnosis of LBP.
  - One member noted that the literature indicates that patients with known anatomic spinal anomalies may not require imaging, although this is not covered in the guidelines. The developers noted they have crafted the measure to align with current guidelines.
- Updated score-level testing results based on the revised specifications were not provided. The developer noted that these data would not be available for analysis until mid-2017.

0052 Use of Imaging Studies for Low Back Pain

- Although the developers did not provide updated data-element level validity testing, they did provide additional information from their 2002 field testing analysis; these data provided insight on the ability to identify patients in the recently-added exclusion categories (i.e., prolonged steroid use, spinal infection, and immunosuppression). The developer stated that according to the administrative data from 2002, the various red-flag conditions specified as exclusions in the measure occur in 0% to 1.9% of LBP episodes (4.9% overall).
- Committee members expressed concern that a significant number of patients with trauma or neurologic impairment are not being captured using administrative claims data. The developer responded that the testing was performed using 2003-2004 data. They suggested the possibility that claims for trauma and neurologic impairment may have improved since then. They also noted a lack of feedback from health plans that the measure is actually missing a lot of trauma cases.
- In their submission materials, the developer presented their expert panel and commenting processes as an assessment of face validity. As part of the discussion, they provided additional explanation of how they believe these processes fulfill NQF's requirement for face validity.
- The Committee's rating of reliability had to depend, to a large extent, one the data element validity testing information provided by the developer (because the developer was unable to update their score-level reliability testing at this time). Therefore, NQF staff allowed for a vote on validity prior to the vote on reliability. Ultimately, the Committee agreed that there was insufficient information provided for validity and did not recommend the measure for endorsement.

STAFF NOTE: In the staff preliminary analysis and during the measure evaluation webinar, NQF staff noted that only percentage agreement statistics were provided to show the level of agreement between administrative codes and medical records and that the results provided were not calculated using the newly-specified measure. However, after further reflection on the additional information that was provided as part of the exclusion analysis (specifically, data on the ability to identify the new exclusions in claims only, in medical records only, or in both), staff determined that these data shed light on questions that are addressed in sensitivity/specificity analysis, even though the developer did not provide actual sensitivity/specificity statistics. Thus, staff no longer considers the data element validity testing presented by the developer to be insufficient. Because we have reversed our previous guidance, NQF will ask the Committee to reconsider and re-vote on Reliability and Validity during the post-comment call. Also, the developer is seeking updated data to present during the post-comment call.

3. Feasibility: H-12; M-7; L-0; I-0

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

STAFF NOTE: Even though this measure did not pass the Validity subcritieria, staff asked the Committee to use the remaining time on the webinar to discuss and vote on the Feasibility and Usability and Use criteria, given that the developer would be providing additional data during the post-comment call.

Rationale:

- All data elements are in defined fields in electronic claims and generated or collected by and used by healthcare personnel during the provision of care. The Committee agreed that the data are readily available and can be captured without undue burden.

4. Usability and Use: H-0; M-18; L-0; I-0

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

Rationale:

- This measure is being used in at least 10 accountability programs, including pay-for-performance programs, accreditation programs, and public reporting.

0052 Use of Imaging Studies for Low Back Pain
• One member asked if there might be unintended consequences of the measure if patients who meet red-flag conditions from the guideline are not being excluded from the measure. The developers acknowledged the possibility, but emphasized their process of obtaining feedback from plans. • The Committee expressed concern about the lack of improvement in performance in the last several years. The developer expressed hope that the Choosing Wisely Campaign would help promote more attention to the issue and drive improvement in performance of the measure.
5. Related and Competing Measures This measure is competing with: • #0514: MRI Lumbar Spine for Low Back Pain (CMS) o Due to differences in the level of analysis and care settings, the Committee will not be asked to select a best in-class measure. o Since the last evaluation, the developers have worked to harmonize the measures, resulting in greater congruence in how low back pain is defined, how cancer exclusions are defined, and in exclusion categories. Areas where the measures are not yet harmonized include the specific imaging modalities captured by the measure and some of the exclusion categories.
Steering Committee Recommendation for Endorsement: **Not Recommended** Rationale: •
6. Public and Member Comment •
7. Consensus Standards Approval Committee (CSAC) Vote: Y-X; N-X
8. Appeals

0514 MRI Lumbar Spine for Low Back Pain
Submission
Description: This measure evaluates the percentage of magnetic resonance imaging (MRI) of the lumbar spine studies for low back pain performed in the outpatient setting where conservative therapy was not attempted prior to the MRI. Antecedent conservative therapy may include claim(s) for physical therapy in the 60 days preceding the lumbar spine MRI, claim(s) for chiropractic evaluation and manipulative treatment in the 60 days preceding the lumbar spine MRI, or claim(s) for evaluation and management at least 28 days but no later than 60 days preceding the lumbar spine MRI. The measure is calculated based on a one-year window of Medicare claims data. The measure has been publicly reported, annually, by the measure steward, the Centers for Medicare & Medicaid Services (CMS), since 2010, as a component of its Hospital Outpatient Quality Reporting (HOQR) Program. **Numerator Statement**: MRI of the lumbar spine studies with a diagnosis of low back pain (from the denominator) without the patient having claims-based evidence of prior antecedent conservative therapy. **Denominator Statement**: The number of MRI of the lumbar spine studies with a diagnosis of low back pain on the imaging claim performed in a hospital outpatient department on Medicare FFS beneficiaries within a 12-month time window. **Exclusions**: Indications for measure exclusion include any patients with the following diagnosis code categories: -Patients with lumbar spine surgery in the 90 days prior to MRI -Cancer (within twelve months prior to MRI procedure) -Congenital spine and spinal cord malformations (within five years prior to MRI procedure) -Inflammatory and autoimmune disorders (within five years prior to MRI procedure) -Infectious conditions (within one year prior to MRI procedure)

0514 MRI Lumbar Spine for Low Back Pain
-Spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage (within five years prior to MRI procedure) -Spinal cord infarction (within one year prior to MRI procedure) -Neoplastic abnormalities (within five years prior to MRI procedure) -Treatment fields for radiation therapy (within five years prior to MRI procedure) -Spinal abnormalities associated with scoliosis (within five years prior to MRI procedure) -Syringohydromyelia (within five years prior to MRI procedure) -Postoperative fluid collections and soft tissue changes (within one year prior to MRI procedure) -Trauma (within 45 days prior to MRI procedure) -IV drug abuse (within twelve months prior to MRI procedure) -Neurologic impairment: (within twelve months prior to MRI procedure) -HIV (within twelve months prior to MRI procedure) -Unspecified immune deficiencies (within twelve months prior to MRI procedure) -Intraspinal abscess (an exclusion diagnosis must be in one of the diagnoses fields on the MRI lumbar spine claim) (Specific CPT codes, ICD-9 codes, and ICD-10 codes for exclusion are included in the value sets for this measure; this detailed list can be found in the Excel workbook provided for criterion S2b.) **Adjustment/Stratification**: No risk adjustment or risk stratification **Level of Analysis:** Facility; Population : Regional and State **Setting of Care:** Emergency Department, Clinician Office/Clinic, Hospital : Hospital, Hospital : Acute Care Facility, Hospital : Critical Care, Imaging Facility, Urgent Care - Ambulatory **Type of Measure**: Process **Data Source**: Claims (Only) **Measure Steward**: Centers for Medicare and Medicaid Services
STEERING COMMITTEE MEETING [1/6/2017] **1. Importance to Measure and Report: <u>The measure meets the Importance criteria</u>** (1a. Evidence, 1b. Performance Gap) 1a. Evidence: **Previous Evidence Evaluation Accepted**; 1b. Performance Gap: **H-3**; **M-10**; **L-0**; **I-0** <u>Rationale</u>: • The developer updated the evidence to include the 2015 American College of Radiology (ACR) Appropriateness Criteria: Low Back Pain. The Committee agreed that this was an appropriate update to the evidence and there was no need to re-discuss and re-vote on the evidence sub-criterion. • To demonstrate opportunity for improvement, the developer provided an analysis of Medicare fee-for-service (FFS) claims data that indicates variation in the use of inappropriate MRI lumbar spine studies. Performance rates for July 2104 to June 2015 averaged 39.5% and ranged from 14.9% to 64.8% (NOTE: a lower rate is better). • Committee members noted that the performance gap data actually demonstrated a *decrease* in performance (from 32.5% in 2009 to 39.5% in 2014-2015). The developer indicated that this could be a result of a change in data sources that were used to compute performance scores. The developer also noted that changes in specifications over time make it difficult to interpret changes in performance across time (specifically, expanding the exclusions would decrease the measure denominator, but would not uniformly affect the measure result). • 2013 data presented by the developer showed that beneficiary age, gender, and race, as well as facility characteristics (i.e., number of beds, urban/rural locality, teaching status) were significantly associated with the rate of inappropriate MRI lumbar spine studies.
2. Scientific Acceptability of Measure Properties: <u>The measure does not meet the Scientific Acceptability</u>

0514 MRI Lumbar Spine for Low Back Pain

criteria

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

2a. Reliability: **H-0**; **M-8**; **L-5**; **I-0** 2b. Validity: **H-0**; **M-3**; **L-9**; **I-1**

Rationale:

- Committee members had several questions and concerns about the measure specifications, as follows:
    - The measure is specified for Medicare Fee-for-Service beneficiaries. However, "elderly individuals" is one of the red-flag conditions in the Appropriate Use guideline, indicating that imaging for the patients presenting with LBP may be appropriate. The developer interpreted the guideline as indicating that "elderly" should not be an independent indicator for imaging; however, some Committee members disagreed with this interpretation.
    - The measure uses evaluation and management (E&M) visits as a proxy for antecedent conservative care (in addition to claims for physical therapy or chiropractic visits). In general, the Committee agreed that the E&M visits are a reasonable proxy for some kinds of antecedent therapy, but questioned whether they would capture other types of antecedent therapy such as telephone encounters. Members noted that some types of antecedent conservative care (e.g., NSAIDs, Tylenol, massage therapy, acupuncture) cannot be captured in claims data.
    - Members questioned several of the look-back periods for some of the exclusions (e.g., 90 days for spine surgery, 12 months for cancer; 5 years for congenital spine and spinal cord malformations). For congenital malformations, the developer clarified that the 5-year look-back was mainly because of lack of access to historical data.
    - Committee members expressed concern that specific codes for neurological impairment, specifically those for which the evidence supports appropriate use of MRI, are not adequately captured in this measure. The developer agreed to look into the coding, but also noted that the red flag conditions often occur in tandem, meaning individual patients often are excluded from the measure due to several of the existing measure exclusions. Committee members noted that sciatica radiculopathy, typically does not present with other red-flag conditions.
- The Committee expressed confusion about what changes, if any, have been made to the measure since the 2014 evaluation. Although the developers described the various analysis they performed (e.g., quantitative and qualitative evaluation of the look-back periods for several of the measure exclusions), it is still not clear if or how the measure has been revised. Some of the confusion dates back to the 2014 evaluation, when the developer had actually added several exclusions to the measure that were not apparent in the submission materials considered by the Committee.
- The developers presented updated score-level signal-to-noise reliability testing using 2013 data. Reliability scores from this analysis ranged from 22.4% to 86.6%, with a median reliability score of 44.9%. The median value was well below 0.7, which is often used as a rule-of-thumb minimal acceptable value, and lower than the 53.1% found in previous testing. The developers also provided, a couple of days prior to the evaluation webinar, another set of testing results. This new testing used a split-sample (or "test-retest") approach to compare agreement in performance across hospitals. The intraclass correlation coefficient from this analysis was 0.59, which can be interpreted as moderate agreement (i.e., there is moderate consistency in performance within facilities).
- The developers assessed the face validity of the measure score by surveying an 11-member Technical Expert Panel (TEP). They asked the TEP members to indicate whether the measure captures the most appropriate and prevalent types of antecedent conservative therapy available through claims data (8 of 11 said yes) and to indicate their agreement as to whether the measure helps assess the inappropriate use of MRI lumbar-spine tests (9 of 11 agreed or strongly agreed).
- The developer clarified that the intent of the measure is not to drive measure results to zero, but to decrease the number of orders for MRI on presentation of LBP and to reduce variation between facilities in inappropriate MRIs.
- After much discussion, the Committee agreed that the measure did not pass the validity subcriterion

0514 MRI Lumbar Spine for Low Back Pain
and did not recommend the measure for endorsement.
3. Feasibility: H-X; M-X; L-X; I-X *(3a. Clinical data generated during care delivery; 3b. Electronic sources; 3c.Susceptibility to inaccuracies/ unintended consequences identified 3d. Data collection strategy can be implemented)* Rationale: •
4. Usability and Use: H-X; M-X; L-X; I-X *(Used and useful to the intended audiences for 4a. Accountability and Transparency; 4b. Improvement; and 4c. Benefits outweigh evidence of unintended consequences)* Rationale: •
5. Related and Competing Measures This measure is competing with: • #0052: Use of Imagine Studies for Low Back Pain (NCQA) o Due to differences in the level of analysis and care settings, the Committee will not be asked to select a best in-class measure. o Since the last evaluation, the developers have worked to harmonize the measures, resulting in greater congruence in how low back pain is defined, how cancer exclusions are defined, and in exclusion categories. Areas where the measures are not yet harmonized include the specific imaging modalities captured by the measure and some of the exclusion categories.
Steering Committee Recommendation for Endorsement: **Not Recommended** Rationale: •
6. Public and Member Comment •
7. Consensus Standards Approval Committee (CSAC) Vote: Y-X; N-X
8. Appeals

On December 12, 2016, the National Committee for Quality Assurance (NCQA) presented *Use of Imaging Studies for Low Back Pain* (NQF #0052) to an ad-hoc meeting of the National Quality Forum's (NQF's) Musculoskeletal Standing Committee. During the meeting, NCQA presented several changes to the measure's denominator and exclusions meant to improve the face validity of the measure.

While the Standing Committee expressed support for the changes, they requested additional information demonstrating the impact of the changes on the measure's reliability and validity. Unfortunately, NCQA will not be able to assess this until we receive health plan submissions in June and have an opportunity to perform analysis in July, 2017. However, in response to the Standing Committee's interest in understanding the potential impact of the changes, NCQA reached out to several plans in order to obtain comparative data. To date, two plans have kindly provided data.

These two plans, covering two regions, provided data demonstrating their performance using the prior measure specification and 2015 data and their performance using the revised specification and 2016 data (Table 1). Both plans submitted data for their Medicaid line of business.

Table 1. Comparison of plan performance using 2015 and 2016 data

	Plan A		Plan B	
	Prior measure (2015 data)	Revised measure (2016 data)	Prior measure (2015 data)	Revised measure (2016 data)
Eligible population	2,069	2,785	1,685	2,205
Number of exclusions	862	484	654	369
Number of numerator events	525	682	404	543
Performance rate	75%	76%	76%	75%

The plans also applied the prior and revised specifications to their 2016 data (Table 2) to demonstrate how the measure changes impacted their rate using the same data set.

Table 2. Comparison of plan performance using 2016 data

	Plan A		Plan B	
	Prior measure (2016 data)	Revised measure (2016 data)	Prior measure (2016 data)	Revised measure (2016 data)
Eligible population	2,007	2,785	1,577	2,205
Number of exclusions	615	484	549	369
Number of numerator events	484	682	349	543
Performance rate	76%	76%	78%	75%

As expected, the eligible population was larger for each plan using the revised specification (NCQA expanded the denominator by adding physical therapy and telehealth visits to the types of visits that make patients eligible for the denominator).

Regarding exclusions, we shortened the timeframe for the recent trauma exclusion from 12 months to three months; this change likely contributed to the smaller number of exclusions observed with the revised specification. Also, as part of the revisions, we expanded the types of conditions which would exclude patients from the denominator; however, these conditions are rare and therefore, they did not have a significant impact on the number of exclusions.

Despite these changes to the denominator and exclusions, the measure performance stayed relatively stable for the two plans.

We will have a better understanding of the impact of these changes when the plans submit their data to NCQA in June, 2017.

To: National Quality Forum Musculoskeletal Standing Committee

From: Vinitha Meyyur [Centers for Medicare & Medicaid Services]
Lori Geary and Arjun Venkatesh [Yale-New Haven Health Services Corporation/Center for Outcomes Research and Evaluation]
Charlie Bruetman and Colleen McKiernan [The Lewin Group]

Date: Friday, March 03, 2017

Subject: Request for Reconsideration of NQF #0514

I. Purpose

The Centers for Medicare & Medicaid Services (CMS) is requesting a reconsideration of the National Quality Forum (NQF) Musculoskeletal Standing Committee's decision not to recommend NQF #0514,[1] MRI Lumbar Spine for Low Back Pain, for continued endorsement. During the January 6 review webinar, NQF #0514 did not pass the *Validity* criterion. The purpose of this memorandum is to provide additional information and rationale to demonstrate that this measure meets the minimum requirements to pass the *Validity* criterion, as defined by NQF, in support of its continued endorsement.

II. NQF #0514: Validity Criterion

NQF guidance evaluates *Validity* based on three sub-criteria: measure specifications, validity testing, and threats to validity. This section describes ways in which NQF #0514 meets each sub-criterion, based on NQF standards.

Measure Specifications

During the January 6 review webinar, Standing Committee members asked what changes had been made to the specifications based on the NQF #0514's 2014 review; in 2014, however, the Consensus Standards Advisory Committee (CSAC) did not recommend any updates to the specifications. Instead, the CSAC requested that the Committee reexamine the measure based on its previous evaluation of the measure's exclusions.

As part of annual measure maintenance activities, CORE and Lewin, on behalf of CMS, have made several changes and updates to the specifications for NQF #0514 to ensure their alignment with updated and newly released clinical practice guidelines. CORE and Lewin also engaged in extensive harmonization activities with the National Committee for Quality Assurance (NCQA), stewards of a related measure for low back imaging, and reviewed all specification changes with a multidisciplinary technical expert panel (TEP), composed of external stakeholders.[2]

---

[1] NQF #0514 was originally endorsed by the Outpatient Imaging Efficiency Steering Committee in October 2008.

[2] The technical expert panel for NQF #0514 is a multistakeholder group of experts that include insurers/purchasers, clinicians, hospital management or administration, patients/patient advocates, and caregivers.

Updates to the specifications since the Standing Committee's 2014 review of NQF #0514 include the addition of congenital spine/spinal cord malformations, inflammatory and autoimmune disorders, infectious conditions, spinal vascular malformations, spinal cord infarctions, effects from radiation, spinal abnormalities associated with scoliosis, syringohydromyelia, and postoperative fluid collections/soft tissue changes, all of which were added to the measure's list of exclusions.[3] CMS has also harmonized NQF #0514's cancer exclusion definition with that of NCQA's measure, by adding ICD codes for neuroendocrine tumors and patients with a personal history of cancer, and has ensured the definitions of low back pain align across both measures.

During the review webinar, select Standing Committee members stated that the specifications for NQF #0514 do not account for several disease states (i.e., sciatica and radicular pain, and degenerative conditions); a subset of diagnosis codes for all three conditions, focusing on the low back and pelvis, are included in the definition of uncomplicated low back pain (the denominator population). Frequently, the appropriate first-line treatment for acute presentation of sciatica/radiculopathy and degenerative conditions that are not associated with other symptoms is attempting antecedent therapy, including at-home intervention (e.g., NSAIDs, rest, or heat/ice packs); clinicians may then order diagnostic imaging, such as a lumbar spine MRI, if the pain associated with these conditions does not improve over time. Immediate imaging for these diagnoses may be indicated, however, if a patient with radicular pain or a degenerative condition presents concomitantly with other symptoms that suggest an underlying etiology for the diagnosis. **We have not received feedback from the measure's TEP during its annual review, nor from external stakeholders, that suggests these three diagnoses should be excluded from the measure's denominator**. We are, however, open to the Committee's feedback and will consider it as we continue to refine the measure during future annual updates.

Select members of the Standing Committee also stated that NQF #0514 should exclude cases of disk herniation from the measure specifications. **The measure's TEP has not suggested adding disk herniation as <u>either</u> an inclusion or an exclusion term.** Disk herniation is also not a reason for immediate imaging based on the ACR *Low Back Pain* Appropriateness Criteria®.

Finally, while the Committee did raise concerns about the ability to capture alternative forms of antecedent conservative therapy (e.g., NSAIDs, massage therapy, and other forms of self-management) using administrative claims data, the measure specifications include a 'treatment window,' a 28- to 60-day period in which patients or providers can attempt management efforts not capturable using OPPS claims preceding a lumbar spine MRI; this window serves as a proxy for those interventions for which claims cannot be submitted to OPPS and helps to identify quick-trigger orders. The 28-day timeframe aligns with guidance from ACR, suggesting clinicians wait at least four weeks before ordering diagnostic imaging for uncomplicated low back pain. In addition to physical

[3] The exclusions listed here were not included in the measure specifications submitted to NQF in March 2014; however, they had been added by the August 2014 in-person meeting at which NQF #0514 was discussed with the Standing Committee. At that time, these new exclusions were presented verbally to the Committee and discussed along with the other, previously added measure exclusions.

therapy or chiropractory, patients with an evaluation and management claim from 28 to 60 days before their MRI are excluded from the measure's numerator.

We believe that the current specifications exceed NQF's minimum requirements for this sub-criterion and demonstrate the measure's validity.

Validity Testing

The NQF *Measure Evaluation Criteria and Guidance* states that "[f]ace 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."[4] Because of the format of the review webinar and the volume of feedback provided by Standing Committee members during the meeting, CORE and Lewin did not have the time or opportunity to detail the rigorous face validity testing process undertaken by CORE and Lewin, or for review of the NQF scoring guidance that should be applied to measures with face validity testing.

Based on NQF guidance, and consistent with many other NQF-endorsed quality measures, CORE and Lewin conducted a systematic assessment of face validity for NQF #0514 by soliciting feedback from members of our TEP. Of the 11 TEP members that responded to the survey, the majority of respondents indicated that the measure has ***strong face validity***. More specifically, to address feedback from Standing Committee members: 8 of 11 TEP members stated that NQF #0514 captured the most appropriate and prevalent types of antecedent therapy available through claims data (1 TEP member responded "Not Sure or Do Not Know"). Nine of the 11 members either agreed or strongly agreed that the measure helps to assess the inappropriate use of MRI lumbar-spine tests (2 TEP members responded "Undecided" or "Do Not Know or Not Applicable").

Our face validity results are robust and meet NQF's standard for a moderate *Validity* rating, as they "[f]ound substantial agreement that the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality."[2]

Threats to Validity

CMS, CORE, and Lewin believe concerns previously raised by the Standing Committee about the measure's threats to validity have been mitigated based on updates to relevant clinical practice guidelines published since NQF #0514's 2014 review.

While the majority of exclusions for NQF #0514 are included in the ACR *Low Back Pain* Appropriateness Criteria®, the measure does capture additional exclusions based on evidence from other clinical practice guidelines.[5] During the January 6 review webinar, the Standing Committee did not raise concerns about the evidence base for this measure and elected ***not*** to re-vote on the *Evidence* criterion. On behalf of CMS, CORE and Lewin have performed rigorous analytic testing for each exclusion to ensure its correct specification; the exclusions were also extensively vetted by

---

[4] National Quality Forum (NQF). 2016 Measure Evaluation Criteria and Guidance. Washington, DC: NQF; 2016. Available at http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=83123. Accessed February 2017.

[5] Please see Appendix A for a list of exclusions and the sources from which they are drawn.

members of the measure's TEP (both through face validity testing and when reviewing results from the exclusions' quantitative evaluation).

Since NQF #0514's 2014 review by this Standing Committee, CMS, CORE, and Lewin performed a series of analyses to respond to Committee feedback. For example, one concern raised during the measure's 2014 review, and briefly discussed on January 6, referenced the look-back period for lumbar spine surgeries, a current measure exclusion for which CMS looks back 90 days. Detailed analyses on this look-back period revealed no significant impact on hospital scoring (for 90 days vs. 1 year vs. 3 years); the current 90-day look back also was supported by the measure's TEP.

In addition to the look-back period for lumbar spine surgeries, members of the Standing Committee were concerned that the look-back window for the cancer and congenital malformation exclusions were too short. This concern was not raised during the Standing Committee's 2014 review of NQF #0514; because no more than five years of claims data is available for use in a CMS claims-based quality measure, we are limited to look backs of five years or less.

During the review webinar for NQF #0514, select Standing Committee members were concerned that older age was not listed as an independent exclusion for the measure, as it appears as a subset of a red-flag condition in the ACR *Low Back Pain* Appropriateness Criteria®. While the Criteria® lists "elderly individual" as a reason for imaging, under variant #2, we believe that old age alone, without the presence of additional symptoms (captured by existing measure exclusions) suggestive of an underlying etiology, is not sufficient to warrant immediate imaging. The Criteria's® literature review states that, "[a] recent study found no statistically significant difference in primary outcome after 1 year for older adults who had spine imaging within 6 weeks after an initial visit for care for LBP versus similar patients who did not undergo early imaging; thus, this panel does not include age older than 50 as an independent red flag."[6] While we agree that age can serve as a proxy for other red-flag conditions, we have already explicitly excluded these conditions from the specifications for NQF #0514. Additionally, on behalf of CMS, CORE and Lewin contacted ACR to review the specifications for NQF #0514 alongside evidence for the *Low Back Pain* Appropriateness Criteria®. Feedback from the ACR Appropriateness Criteria Panel stated that imaging within the elderly remains a controversial topic, as there is less research into outcomes for this population. ACR Panel members also noted that there is not consensus in the literature on the age threshold for defining an 'elderly individual,' as a number of factors (including lifestyle choices and genetics) contribute to a person's overall health. Although members of the Standing Committee suggest NQF #0514 exclude patients over age 70 years, we continue to include older adults, given that the current measure exclusions capture many of the reasons for imaging within the older adult population. We are, however, open to discussing additional red-flag conditions with the Standing Committee for which early imaging may be appropriate within this population.

---

[6] Patel ND, Broderick DF, Burns J, Deshmukh TK, Fries IB, Harvey HB, Holly L, Hunt CH, Jagadeesan BD, Kennedy TA, O'Toole JE, Perlmutter JS, Policeni B, Rosenow JM, Shroeder JW, Whitehead MT, Cornelius RS, Corey AS, Expert Panel on Neurologic Imaging. ACR Appropriateness Criteria® low back pain. Reston (VA): American College of Radiology (ACR); 2015.

Based on the information presented above, **CMS has minimized the threats to validity for NQF #0514, by obtaining feedback from stakeholders and performing rigorous testing, and believe the minimum standards for the *Threats to Validity* criterion have been met.**

III. Conclusion

Based on NQF's *Measure Evaluation Criteria and Guidance*,[2] we believe that NQF #0514 aligns with the moderate validity recommendation from algorithm #3 (Guidance for Evaluating Validity), as it has received in prior evaluations for endorsement. The measure specifications are aligned with the most updated clinical practice guidelines and have strong face validity; additionally, measure testing confirms that threats to validity have been addressed by the exclusion of red-flag conditions. NQF #0514 also passed the *Importance* and *Reliability* criteria during endorsement maintenance review. As one Standing Committee member stated during the review webinar, there will always be exceptions in health care, and, as long as the rate of exceptions is low, performance scores will not be impacted and the measure serves its purpose; **we believe that, as currently specified, the measure addresses the broader patterns of care**.

CMS, CORE, and Lewin agree the web-based format of the January 6 review webinar introduced challenges that prevented a facilitated assessment of NQF #0514 against NQF criteria. The meeting structure made it difficult for the Standing Committee and NQF staff to engage in a productive dialogue and for CMS, CORE, and Lewin to answer Committee member questions specific to each validity sub-criterion. Accordingly, NQF's evaluation algorithm could not easily be followed, step-by-step, as normally occurs during in-person review, resulting in many important details and supporting materials not being readily available to Standing Committee members participating on the call.

Reducing variation and improving the efficiency of MRI imaging for patients with low back pain is nationally recognized as an important target for quality improvement. NQF #0514 addresses this national priority and has been successfully incorporated into a CMS quality reporting program for six years without evidence of unintended consequences.

CMS submits this clarifying memo and requests the Musculoskeletal Standing Committee's reconsideration of the *Validity* criterion for NQF #0514 to maintain its continued NQF endorsement.

Appendix A: NQF #0514 Exclusion Crosswalk

Exclusion	Look-Back Period	Year Added	Is this an ACR Criteria?	ACR Rating	Source (Non-ACR Recommendations)
Cancer	12 Months	2008	Yes	Without Contrast - 7 Without and With Contrast - 8	-
Congenital Spine and Spinal Cord Malformations	60 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
Infectious Conditions	12 Months	2015	Yes	Without Contrast - 7 Without and With Contrast - 8	-
			No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
Inflammatory and Autoimmune Disorders	60 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
			No	-	Diagnostic imaging practice guidelines for musculoskeletal complaints in adults-an evidence-based approach-part 3: spinal disorders. J Manipulative Physiol Ther Bussieres AE, Taylor JA, Peterson C. 2008 Jan;31(1):33-

Exclusion	Look-Back Period	Year Added	Is this an ACR Criteria?	ACR Rating	Source (Non-ACR Recommendations)
					88.
Intraspinal Abscess	On Claim	2011	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
HIV	12 Months	2011	Yes	Without Contrast - 7 Without and With Contrast - 8	-
			No	-	Michigan Quality Improvement Consortium, 2010. Management of acute low back pain.
			No	-	University of Michigan Health System. Acute low back pain. Ann Arbor (MI): University of Michigan Health System. 2010.
			No	-	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
			No	-	American Academy of Neurology. Practice parameters: Magnetic resonance imaging in the evaluation of low back syndrome (Summary statement). Neurology. 2013;

Exclusion	Look-Back Period	Year Added	Is this an ACR Criteria?	ACR Rating	Source (Non-ACR Recommendations)
					44:767-770
			No	-	Work Loss Data Institute. Low back - lumbar & thoracic (acute & chronic). Corpus Christi (TX): Work Loss Data Institute. 2011
IV Drug Abuse	12 Months	2008	Yes	Not Rated	-
			No	-	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
			No	-	Michigan Quality Improvement Consortium. Management of acute low back pain. Southfield (MI): Michigan Quality Improvement Consortium. 2012.
			No	-	University of Michigan Health System. Acute low back pain. Ann Arbor (MI): University of Michigan Health System. 2010.
Lumbar Spine Surgery	3 Months	2013	Yes	Without Contrast - 8 Without and With Contrast - 5	-

Exclusion	Look-Back Period	Year Added	Is this an ACR Criteria?	ACR Rating	Source (Non-ACR Recommendations)
Neoplastic Abnormalities	60 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
			No	-	Diagnostic imaging practice guidelines for musculoskeletal complaints in adults-an evidence-based approach-part 3: spinal disorders. J Manipulative Physiol Ther Bussieres AE, Taylor JA, Peterson C. 2008 Jan;31(1):33-88.
Neurologic Impairment	12 Months	2008	Yes	Without Contrast - 9 Without and With Contrast - 8	-
Postoperative Fluid Collections and Soft Tissue Changes	12 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
Spinal Abnormalities Associated with Scoliosis	60 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
			No	-	Diagnostic imaging practice guidelines for musculoskeletal complaints in adults-an evidence-based approach-part 3: spinal disorders. J Manipulative Physiol Ther

Exclusion	Look-Back Period	Year Added	Is this an ACR Criteria?	ACR Rating	Source (Non-ACR Recommendations)
					Bussieres AE, Taylor JA, Peterson C. 2008 Jan;31(1):33-88.
Spinal Cord Infarction	12 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
Spinal Vascular Malformations and/or the Cause of Occult SAH	60 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
Syringohydromyelia	60 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.
Trauma	45 Days	2008	ACR Low Back Pain	Without Contrast - 7	-
Treatment Fields for Radiation	60 Months	2015	No	-	Parameter for the Performance of Magnetic Resonance Imaging (MRI) of the adult spine. The American College of Radiology and the American Society of Neuradiology (ACR/ASNR) [online publication] Reston, VA; 2012.

Exclusion	Look-Back Period	Year Added	Is this an ACR Criteria?	ACR Rating	Source (Non-ACR Recommendations)
Unspecified Immune Deficiencies	12 Months	2011	Yes	Without Contrast - 7 Without and With Contrast - 8	-
			No	-	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
			No	-	Michigan Quality Improvement Consortium. Management of acute low back pain. Southfield (MI): Michigan Quality Improvement Consortium. 2012.
			No	-	University of Michigan Health System. Acute low back pain. Ann Arbor (MI): University of Michigan Health System. 2010.

Side-By-Side Comparisons

NQF#	0052 - NCQA	0514 - CMS
Description	The percentage of patients with a primary diagnosis of low back pain who did not have an **imaging study** (plain X-ray, MRI, CT scan) within 28 days of diagnosis	The percentage of MRI **of the lumbar spine** studies for low back pain performed in the outpatient setting where conservative therapy was not attempted prior to the MRI
Better quality	Higher score	Lower score
Data Source	Administrative Claims	Administrative Claims
Level of Analysis	Health Plan, Integrated Delivery System	Facility, Region, State
Setting	Clinician Office/Clinic, Emergency Department, Ambulatory Urgent Care	Clinician Office/Clinic, Emergency Department, Hospital-Acute/Critical Care Facility, Imaging Facility, Ambulatory Urgent Care
Numerator	**X-Ray, CT, MRI** within 28 days of LBP dx	**MRI** without evidence of prior antecedent conservative therapy (PT/chiropractic treatment in 60 prior, E&M visit between 28-60 prior
Denominator	Patients ages **18-50** with primary dx of uncomplicated LBP (claims from outpatient visit, observation visit, ED visit, osteopathic/chiropractic treatment, PT visit, telehealth visit)	MRIs of **Medicare FFS beneficiaries** with LBP dx (hospital outpatient only)

Side-By-Side Comparisons

NQF#	0052 - NCQA	0514 - CMS
Exclusions	• Recent diagnosis (6 months prior) of uncomplicated low back pain • Cancer – history of to 28 days after IESD • Trauma -3 months prior to IESD to 28 days after IESD • Recent IV drug abuse –12 months prior to IESD to 28 days after IESD • Neurologic impairment–12 months prior to IESD to 28 days after IESD • HIV— history of to 28 days after IESD • Spinal infection–12 months prior to IESD to 28 days after IESD • Major organ transplant–history of to 28 days after IESD • Prolonged use (90 days) of corticosteroids–12 months prior to IESD and including IESD • Hospice enrollees???	• Lumbar spine surgery within 90 days prior • Cancer within 12 months prior • Neoplastic abnormalities within 5 years prior • Trauma within 45 days prior • IV drug abuse within 12 months prior • Neurologic impairment within 12 months prior • HIV within 12 months prior • Unspecified immune deficiencies within 12 months prior • Inflammatory and autoimmune disorders within 5 years prior • Infectious conditions within 1 year prior • Congenital spine and spinal cord malformations within 5 years prior • Spinal vascular malformations and/or the cause of occult subarachnoid hemorrhage within 5 years prior • Spinal cord infarction within 1 year prior • Treatment fields for radiation therapy within 5 years prior • Spinal abnormalities associated with scoliosis within 5 years prior • Syringohydromyelia within 5 years prior • Postoperative fluid collections and soft tissue changes within 1 year prior • Intraspinal abscess