

Accepted Manuscript

Title: Variability in diagnostic error rates of ten MRI centers performing lumbar spine MRI exams on the same patient within a three week period

Author: Richard Herzog, Daniel R Elgort, Adam E Flanders, Peter J. Moley

PII: S1529-9430(16)31093-2

DOI: <http://dx.doi.org/doi: 10.1016/j.spinee.2016.11.009>

Reference: SPINEE 57207

To appear in: *The Spine Journal*

Received date: 12-7-2016

Revised date: 22-9-2016

Accepted date: 14-11-2016



Please cite this article as: Richard Herzog, Daniel R Elgort, Adam E Flanders, Peter J. Moley, Variability in diagnostic error rates of ten MRI centers performing lumbar spine MRI exams on the same patient within a three week period, *The Spine Journal* (2016), <http://dx.doi.org/doi: 10.1016/j.spinee.2016.11.009>.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Variability in Diagnostic Error Rates of Ten MRI Centers Performing Lumbar Spine MRI Exams on the Same Patient Within a Three Week Period

Author List:

1. Richard Herzog, MD, FACR, Hospital for Special Surgery, Spremo Quality Research Institute
2. Daniel R Elgort, PhD, Spremo Quality Research Institute
3. Adam E Flanders, MD, Thomas Jefferson University Hospital
4. Peter J. Moley, MD, Hospital for Special Surgery

Corresponding author:

Richard Herzog, MD, FACR

MRI Division, Department of Radiology and Imaging, Hospital for Special Surgery, 535 E. 70th St, New York, NY 10021, USA.

E-mail address: herzogr@hss.edu

Acknowledgments:

We thank Natasha Irani and Orlando Castaneda for their contributions to the study, including data acquisition, data analysis, data visualization, and literature review.

Author disclosures:

The study was approved by the Hospital for Special Surgery IRB.

The study was funded by the Spremo Quality Research Institute.

Abstract

Background Context:

In today's healthcare climate, Magnetic Resonance Imaging (MRI) is often perceived as a commodity -- a service where there are no meaningful differences in quality and thus an area in which patients can be advised to select a provider based on price and convenience alone. If this prevailing view is correct, then a patient should expect to receive the same radiological diagnosis regardless of which imaging center he or she visits or which radiologist reviews the examination. Based on their extensive clinical experience, the authors believe that this assumption is not correct and that it can negatively impact patient care, outcomes and costs.

Purpose:

This study is designed to test the authors' hypothesis that the radiologist's reports from multiple imaging centers performing a lumbar MRI exam on the same patient over a short period of time will have (1) marked variability in interpretive findings and (2) a broad range of interpretive errors.

Study Design:

A prospective observational study comparing the interpretive findings reported for one patient scanned at 10 different MRI centers over a period of three weeks to each other and to reference MRI exams performed immediately preceding and following the 10 MRI exams.

Patient Sample:

A 63-year old female with a history of low back pain and right L5 radicular symptoms.

Methods:

The complete set of interpretive findings from the 10 study MRI exams were tabulated and compared for variability and errors. Two of the authors, both subspecialist spine radiologists from different institutions, independently reviewed the reference exams and then came to a final diagnosis by consensus. Errors of interpretation in the study exams were considered present if a finding present or not present in the study exam's report was not present in the reference exams.

Outcome Measures:

Variability was quantified using percent agreement rates and Fleiss' Kappa statistic. Interpretive errors were quantified using true positive counts, false positive counts, false negative counts, true positive rate (sensitivity), and false negative rate (miss rate).

Results:

Across all 10 study exams there were 49 distinct findings reported related to the presence of a distinct pathology at a specific motion segment. Zero interpretive findings were reported in all 10

study exams and only one finding was reported in nine out of 10 study exams. 32.7% of the interpretive findings appeared only once across all 10 of the study exams' reports. A global Fleiss' Kappa statistic, computed across all reported findings, was 0.20 ± 0.06 , indicating poor overall agreement on interpretive findings. The average interpretive error count in the study exams was 12.5 ± 3.2 (both false positives and false negatives). The average false negative count per exam was 10.9 ± 2.9 out of 25 and the average false positive count was 1.6 ± 0.9 , which corresponds to an average true positive rate (sensitivity) of $56.4\% \pm 11.7$ and miss rate of $43.6\% \pm 11.7$.

Conclusions:

This study found marked variability in the reported interpretive findings and a high prevalence of interpretive errors in the radiologists' reports of an MRI exam of the lumbar spine performed on the same patient at 10 different MRI centers over a short time period. As a result, the authors' conclude that where a patient obtains their MRI exam and which radiologist interprets the exam may have a direct impact on their radiologic diagnosis, subsequent choice of treatment and clinical outcome.

Key Words: MRI lumbar spine, MRI diagnostic variability, MRI diagnostic error rates, MRI diagnostic accuracy, MRI quality

1 **Introduction:**

2 In the clinical evaluation of a patient with back or leg pain unresponsive to conservative
3 measures, clinicians may order an MRI exam to assist in explaining the patient's symptoms in
4 order to determine whether or not modification of the patient's therapy is required, including
5 referral for interventional pain management and/or surgical evaluation. Moreover, the results of
6 MRI exams play a central role when payers are reviewing whether or not to approve a
7 recommended treatment. Therefore, an accurate diagnosis is paramount to timely and correct
8 treatment. Several studies provide information as to the variability of interpretation of
9 radiological exams, including MRI exams of the lumbar spine, and the importance of
10 nomenclature when communicating radiologic findings [1 - 11]. However, these studies provide
11 no information as to the variability and quality of interpretation of all MRI findings in a single
12 patient imaged at different imaging centers. The authors believe that their study presented here is
13 the first of its kind and provides critically important and novel insights into the variability and
14 diagnostic performance between MRI exams.

17 **Method:**

18 The study subject was a 63-year old female with a history of low back pain and right L5
19 radicular symptoms. Her pain radiated down to the anterolateral side of her right leg. On
20 examination, she had mild weakness in right ankle dorsiflexion 4+/5 and right great toe
21 extension 4+/5, reflexes were diminished, but symmetrical bilaterally (1/4), and she had a
22 positive dural tension (Seated Slump) sign on the right. After IRB approval, the subject
23 underwent 12 MRI exams of the lumbar spine. Ten exams were performed at 10 different

regional imaging centers over a period of three weeks along with two reference MRI exams performed at one of the author's institutions immediately preceding and following the 10 MRI exams. The reference exams were performed on a closed 1.5T MRI system and included the following sequences: Spin-echo T1, spin-echo T2, and STIR sagittal sequences (all sequences were acquired with a slice thickness 3.5 mm/no gap and a minimum of 24 slices); three stacked overlapping spin-echo T2 axial sequences were acquired perpendicular to the central canal and parallel to a disc space: one from T12 to L3 (parallel to the L2-3 disc space), the second from L3 to S1 (parallel to the L4-5 disc space), and the third from the top of L5 to the bottom of S1 (parallel to the L5-S1 disc space); and a spin-echo T2 coronal sequence acquired parallel to the posterior cortex of L4 and included the majority of the lumbar vertebral bodies, the entire lumbar central canal and all of the lumbar posterior elements. The ten study centers performed their routine MRI exam.

The study centers were selected by their location within or close proximity to New York City, for the convenience of the study patient, and for a range of MRI equipment. The equipment used across the 10 study centers included one open 0.3T, one stand-up 0.6T, seven closed 1.5T and one closed 3.0T MRI system. The authors verified that all study centers had valid accreditation from the American College of Radiology (ACR), including the Spine MRI module [12], at the time of this study. The 10 MRI centers were blinded to their participation in the study and evaluated the subject as a routine patient. The subject presented to each MRI center with the same prescription completed by an orthopedic surgeon unaffiliated with either author's institution. The prescription stated that the patient had back and leg pain. If requested verbally, or

1 as part of an intake questionnaire, the subject provided the same history of back pain and leg pain
2 to each MRI center. The subject did not reveal that she was participating in a clinical study.

3
4 Following completion of the 10 study MRI exams, the MRI reports from these exams were
5 stripped of all information identifying the center, radiologist, and type of equipment used. The
6 reports were then reviewed by one of the authors, a subspecialized spine radiologist, and all the
7 reported findings (appearing in either the Body or Impression sections) were inserted into a
8 single “Study Exam Data Sheet” for cross-exam comparisons. Interpretive variability was then
9 quantified using percent agreement rates and Fleiss’ Kappa statistic.

10
11 Two of the authors, both subspecialist spine radiologists from different institutions and with over
12 25 years of clinical experience, independently reviewed the two reference MRI exams. The only
13 discussion of the two authors prior to independently interpreting the exams was to confirm the
14 grading system for stenosis [9]. Only three minor disagreements in findings related to the
15 severity of neural foraminal stenosis had to be resolved by consensus and the final set of findings
16 was used as the reference findings.

17
18 For the purpose of which reference findings to use for the evaluation of interpretive errors, the
19 authors limited the findings to a subset of findings that were reported in the Spinal Patient
20 Outcomes Research Trial (SPORT) [9, 10]. Specifically, the reference findings were limited to:
21 disc degeneration, disc herniation, spinal stenosis, nerve root involvement, facet degeneration,
22 anterior spondylolisthesis, and vertebral fracture. The diagnosis of a disc herniation was based on
23 detecting a localized or focal displacement of disc material beyond the limits of the intervertebral

disc space [5]. The type of disc herniation, i.e. protrusion, extrusion or sequestered fragment, was not captured since many study exam reports did not make this differentiation. The diagnosis of central canal stenosis was based on the visual assessment of thecal sac cross-sectional area [9]. The area of the thecal sac at the level of the disc space (or the level of the most severe stenosis) was compared to the area of the thecal sac at the level of the pedicles cephalad to the level of stenosis. Stenosis was considered mild if the thecal sac area was reduced by one-third or less, moderate if reduced between one- and two-thirds, and severe if reduced by more than two-thirds. Neural foraminal stenosis was graded by visually assessing the reduction in the area of the neural foramen and was considered mild if the area was reduced by one-third or less, moderate if reduced by between one- and two-thirds, and severe if reduced by more than two-thirds [9]. Nerve root involvement was diagnosed based on the presence of any pathologic process that abutted, impinged, displaced or compressed a nerve root or the presence of an anomalous nerve root.

The reference findings were then compared to the study exam findings collected in the Study Exam Data Sheet to identify interpretive errors. An error in interpretation in a study exam was considered present if there was no mention in the report of a reference finding. Any positive finding reported in a study exam that was not present in the reference findings was also considered an error in interpretation. There were no instances in which a positive finding was reported in a study center exam that was missed by the two independent reviewers during evaluation of the reference exams.

In order to reduce sensitivity related to the lack of accepted standards for the measurement of stenosis, the authors only recorded an error if the grading was over-called or under-called by two grades (e.g. severe was present and only called mild). Similarly, in order to reduce over-reporting errors resulting from variation of nomenclature for degenerative disc disease, the authors accepted all of the following: disc degeneration, disc bulge with reduced T2 nuclear signal intensity, disc desiccation, spondylosis and decreased disc height to indicate disc degeneration.

Interpretive errors were quantified using true positive counts, false positive counts, false negative counts, true positive rate (sensitivity), and false negative rate (miss rate). Accuracy was not used as a statistical metric in this study because silence on any pathology in a report was interpreted as a negative finding, which makes quantifying true negatives problematic.

Results:

Interpretive Variability: There was marked variability in reported findings across the 10 study exams. Across all 10 exams there were 49 distinct findings reported (in either the Body or the Impression section of the MRI reports) related to the presence of a distinct pathology at a specific motion segment. The findings included: vertebral alignment, disc bulge, disc degeneration/desiccation or spondylosis, disc height, disc herniation, stenosis of the central canal, lateral recess and neural foramina, nerve root involvement, endplate degeneration and facet degeneration. Among the noteworthy aspects of this aggregated set of findings is that none of the 49 reported findings were unanimously reported in all 10 study exams and only one of the

findings, the anterior spondylolisthesis present at L5-S1, was reported in nine out of 10 exams. 32.7% of the interpretive findings only appeared once across all 10 reports (**Figure 1**).

The overall level of agreement on the reported findings of the study exams was summarized using Fleiss' Kappa statistic, a standard measure of inter-rater agreement used for data with multiple raters that accounts for the likelihood of agreements due to random chance [13]. The Fleiss' Kappa statistic can have a maximum value of 1.0, indicating perfect agreement among the imaging exams' reports. A Fleiss' Kappa statistic value of zero or less than zero, indicates that the level of agreement is no better than chance. Generally, values above 0.75 are considered to indicate excellent agreement, values between 0.4 to 0.75 are interpreted as intermediate/good agreement, and values below 0.4 are interpreted as poor agreement [13]. The overall Fleiss' Kappa statistic across the 10 exams and all reported interpretive findings was 0.20 ± 0.03 , indicating poor overall agreement on interpretive findings.

To illustrate the variation in the study exams' reported interpretive findings, **Figure 2** depicts how a disc herniation was reported in the 10 exams. The number of exams reporting the presence of a disc herniation at a given motion segment ranged from 70% at L3-4 to 20% at L5-S1; two exams reported a disc herniation at all five motion segments and one exam did not report a disc herniation at any motion segment. The number of study exams reporting thecal sac compression due to a disc herniation ranged from 60% at L1-2 to only one exam reporting thecal sac compression at L4-5. Nerve root involvement due to a disc herniation was reported in 20% of the exams at L2-3, 40% of the exams at L3-4, and 30% of the exams at L4-5. The Fleiss' Kappa

score for agreement on the presence of a disc herniation was -0.02 ± 0.23 across the five motion segments.

Similar variation existed with respect to reporting stenosis in the study exams. The number of study exams reporting the presence of central canal stenosis at a given motion segment ranged from 80% at L3-4 to only one exam reporting central canal stenosis at L1-2. Central canal stenosis was reported at four motion segments in two exams and not present at any motion segment in two exams. The Fleiss' Kappa score for agreement on the presence of central canal stenosis was 0.17 ± 0.32 across the five lumbar motion segments.

Only five out of the 10 exam reports included descriptions of any effect of spinal pathology on nerve roots. The number of study exams reporting the presence of nerve root involvement at a given motion segment ranged from 50% at L3-4 to only one exam reporting nerve root involvement at L5-S1. In four study exams, nerve root involvement was reported at three motion segments and in five study exams, nerve root involvement was not reported as present at any motion segment.

Interpretive Diagnostic Errors:

In addition to the significant variability in reported findings, there was a high rate of interpretive errors across the study exams, based on comparing the study exams to the reference exams. The study exams had the lowest interpretive miss rate, 10%, with respect to the patient's single instance of anterior spondylolisthesis, and the highest miss rate, 72.5%, for the patient's four

instances of nerve root involvement. The interpretive miss rates for all other pathologies ranged from 30% to 47.5% summarized in **Table 1**.

Figure 3 illustrates an example from the reference exam for grading central canal stenosis. At level of the L2 pedicles the area of the thecal sac measures approximately 241 sq. millimeters and at level of the L2-3 disc space the area of the thecal sac measures approximately 67 sq. millimeters. The reduction of the thecal sac is greater than 2/3 and graded as severe stenosis. At L2-3 the central canal stenosis was not reported in four, reported as moderate in five and severe in one study MRI exam (no error of interpretation was assessed for reporting the stenosis as moderate since, as indicated in the Methods section, an error in interpretation was assessed only if the stenosis was misgraded by two grades). The patient's other level of severe central canal stenosis was not reported in two, reported as mild in 3, moderate in 4 and severe in one of the study MRI exams.

Table 2 summarizes the interpretive errors of each study exam report compared to the set of reference findings. The study exams' average interpretive error count was 12.5 ± 3.2 per exam (both false positives and false negatives). The study exams' average false negative count was 10.9 ± 2.9 and their average false positive count was 1.6 ± 0.9 . This corresponds to an average true positive rate (sensitivity) of $56.4\% \pm 11.7$ and false negative rate (miss rate) of $43.6\% \pm 11.7$.

Discussion:

1 This study is the first in which interpretive variability and error rates were assessed across 10
2 complete lumbar MRI exams of the same patient, conducted at 10 unaffiliated imaging centers
3 within a three week period, and interpreted by radiologists who were blinded to their
4 participation in the study. This study identified marked variability in the reported interpretive
5 findings and an alarmingly high number of interpretive errors in the lumbar MRI reports. With
6 respect to variability, no interpretive findings were reported in all 10 study exams and only one
7 finding was reported in nine out of 10 study exams. 32.7% of the interpretive findings only
8 appeared once across all 10 of the study exams' reports. A global Fleiss' Kappa statistic,
9 computed across all reported findings, was 0.20 ± 0.06 , indicating poor overall agreement on
10 interpretive findings. The level of variability across the exams in this study is higher than the
11 variability reported in previous studies in the literature which assessed variability of
12 interpretation of the same set of images for multiple patients and grading a restricted set of
13 pathologies and employing a pre-agreed set of definitions in most studies [7, 8, 9, 10, 11].
14 Quantifying the prevalence and types of interpretive errors in the study exams, there were an
15 average of 12.5 ± 3.2 interpretive errors (both false positives and false negatives) across the 10
16 MRI exams. The high average interpretive miss rate of $43.6\% \pm 11.7$ across the study exams
17 means that important pathologies are routinely under reported. For example, the study exams
18 demonstrated an average miss rate for disc herniation of 47.5%. Similarly, the high false positive
19 rates for specific pathologies indicate that diagnostic findings, such as central canal stenosis, may
20 be routinely overcalled.

21
22 The authors acknowledge that many physicians, in particular spine specialists, are generally able
23 to independently review the MRI exam to verify the reported findings in order to formulate the

1 most appropriate treatment plan including surgical care. But some physicians who provide care
2 in the acute stages of a patient's symptoms (e.g., family practice, internal medicine) are not as
3 well-trained or experienced in reviewing MRI exams. As a result, the initial diagnosis and
4 treatment recommendations may be based on an inaccurate MRI interpretation resulting in
5 incorrect treatment recommendations, delayed recovery and/or poor outcome. Moreover, for
6 patients being considered for less invasive procedures (e.g., interventional pain management or
7 other minimally invasive procedures), there may be an overreliance on the MRI report.
8 Importantly, it would be an omission not to add that the payer community heavily relies upon
9 MRI reports during utilization and authorization review procedures. As a result, an incorrect
10 diagnosis on an MRI has the potential to significantly delay authorization of appropriate care that
11 in turn can negatively impact patient outcomes.

12
13 Several limitations of the current investigation exist due to study design and practical constraints.
14 The first limitation is that since only a single MRI exam was performed at each of the 10 study
15 MRI centers, the results reflect a single radiologist at each study center and may not be reflective
16 of the overall performance of the MRI center. For this reason, the authors were unable to
17 evaluate whether or not these findings were representative of the imaging centers selected for the
18 study or generalizable to other MRI centers. Second, the sample size and geographic distribution
19 of the study centers needed to be restricted in this study for logistical reasons to ensure the
20 centers were accessible to the patient and could all be visited within a short time frame, as well
21 as to respect the limits of the patient's tolerance for repeated exams. Third, the authors selected a
22 set of study centers employing a range of equipment types to reflect the variation present in the
23 regional market of the subject. This distribution may deviate from the true distribution outside of

1 the study area. Moreover, since many of the equipment types were only used for one or two of
2 the exams, there was not sufficient sample sizes to make statistically meaningful conclusions
3 about the impact of MRI scanner type (e.g., 0.3T vs. 1.5T vs 3.0 T) on the variability or
4 interpretive error rates. Fourth, for similar reasons as above, the study was unable to evaluate the
5 correlation between variability or errors and exam cost or other characteristics that may have
6 varied across study centers and radiologists.

7
8 Furthermore, the diagnostic variability and interpretive error rates observed in this study of a
9 single patient may not be fully generalizable to all patient cohorts and pathologies. Using a single
10 patient for the study limits the type and severity of pathology available for comparison. When
11 selecting the single patient for this study an effort was made to recruit a subject with a non-trivial
12 number and range of pathologies present in the lumbar spine, which the authors believed would
13 allow for a more concrete comparison and evaluation of the interpretive performance. As a
14 result, this study design likely had some inherent bias toward detecting more false negative
15 interpretive errors than false positives.

16
17 Due to these limitations, the authors did not attempt to identify or assess the relative importance
18 of factors that explain the observed variability and errors across the 10 study exams. However,
19 potential reasons for the variability in the interpretation of the MRI exams and prevalence of
20 interpretive errors include the degree of specialization of the radiologist interpreting the MRI
21 exam, the type of equipment and imaging sequences used at the study centers and the
22 nomenclature employed by the radiologists to describe and communicate abnormalities detected
23 in the MRI exam. The authors did not attempt to train the radiologists at the 10 study centers on

1 spinal nomenclature, since the study was designed to simulate what is currently occurring in the
2 medical community where there is little agreement on the nomenclature used to describe many
3 spinal pathologies. Moreover, the omission or inclusion of pathological findings may vary based
4 on community standards for a variety of reasons, including but not limited to the opinions of the
5 referring physicians and the interpreting radiologists as to how distinct findings may be
6 contributory to a patient's symptoms. The authors acknowledge that the potential reasons cited
7 for the variability in the interpretation are speculative and additional important factors may also
8 be contributing to the observed variability.

9
10 Notwithstanding the study limitations, these results highlight critical issues and provide some
11 novel insights and perspective. The study centers are representative of segments of the diagnostic
12 MRI market actively treating patients. Even before addressing this study's results regarding
13 interpretive errors, the underlying level of variability across the centers' MRI reports should be
14 cause for concern. The fact that no interpretive finding was reported unanimously by the
15 radiologist at all centers and that one-third of all reported findings only appeared once across all
16 10 study exam reports indicates that there is at best, significant difference in the standards
17 employed by radiologists when deciding what to include in diagnostic reports, and at worst
18 significant prevalence of interpretive errors.

19
20 Based on the variability and interpretative errors identified in this study, further investigation is
21 required to understand the cause of these findings and their impact on the trajectory of patient
22 care, outcomes and costs. Moreover, awareness of the prevalence of errors may benefit
23 providers in circumstances where there is poor correlation between a patient's clinical

1 presentation and the reported MRI findings. Ultimately, it is the authors' opinions that accurate
2 and complete diagnostic information at the onset of an injury or illness is critical to improve the
3 chances for a patient's full recovery. However, reducing diagnostic errors and variability in
4 reported findings will require the development and adoption of systematic mechanisms for
5 measuring diagnostic MRI quality, including error rates. The authors acknowledge that
6 accurately measuring interpretive errors at scale is a significant challenge and that some
7 healthcare providers may be reluctant to adopt such a system due to concerns around exposure of
8 their errors, negative impact on reimbursement, and potential liability. Broad acceptance of the
9 prevalence of errors and their potential impact on care is a critical first step toward a system
10 capable of providing industry-wide, standardized measurement of diagnostic MRI quality.

References

- [1] Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. *Am J Roentgen* 2013;201:611–17. doi:10.2214/AJR.12.10375
- [2] Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. *J Med Imaging and Radiation Oncol* 2012;56:173–78. doi:10.1111/j.1754-9485.2012.02348.x
- [3] Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *RadioGraphics* 2015;35:1668–76
- [4] Diaz S, Ekberg O. The frequency of diagnostic errors in radiologic reports depends on the patient's age. *Acta Radiol* 2010;8:934-38. doi 10.3109/02841851.2010.503192
- [5] Fardon DF, Williams AL, Dohring EJ, Murtagh FR, Rothman SLG, Sze GK. Lumbar disc nomenclature: version 2.0 recommendations of the combined task forces of the North American Spine Society, the American Society of Spine Radiology and the American Society of Neuroradiology. *Spine J* 2014;14:2525–45. <http://dx.doi.org/10.1016/j.spinee.2014.04.022>
- [6] Li Y, Fredrickson V, Resnick D. How should we grade lumbar disc herniation and nerve root compression? A systematic review. *Clin Ortho and Relat Res* 2015;473:1896–1902. doi 10.1007/s11999-014-3674-y
- [7] Fu MC, Buerba RA, Long WD, Blizzard DJ, Lischuk AW, Haims AH, Grauer JN. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. *Spine J* 2014;14:2442–48. <http://dx.doi.org/10.1016/j.spinee.2014.03.010>

- [8] Weber C, Rao V, Gulati S, Kvistad KA, Nygaard OP, Lonne G. Inter- and intraobserver agreement of morphological grading for central lumbar spinal stenosis on magnetic resonance imaging. *Glob Spine J* 2015;5:406–10. [http://dx.doi.org/ 10.1055/s-0035-1551651](http://dx.doi.org/10.1055/s-0035-1551651)
- [9] Lurie JD, Tosteson AN, Tosteson TD, Carragee E, Carrino J, Kaiser J, Sequeiros RTB, Lecomte AR, Grove MR, Blood EA, Pearson LH, Weinstein JN, Herzog R. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. *Spine* 2008;33(14):1605–10. doi:10.1097/BRS.0b013e3181791af3
- [10] Carrino JA, Lurie JD, Tosteson ANA, Tosteson TD, Carragee EJ, Kaiser J, Grove MR, Blood E, Pearson LH, Weinstein JN, Herzog R. Lumbar spine: Reliability of MR imaging findings." *Radiol* 2009;250(1):161-70.
- [11] Speciale AC, Pietrobon R, Urban CW, Richardson WJ, Helms CA, Major N, Enterline D, Hey L, Haglund M, Turner DA. Observer variability in assessing lumbar spinal stenosis severity on magnetic resonance imaging and its relation to cross-sectional spinal canal area. *Spine* 2002;27:1082–186.
- [12] American College of Radiology. MRI accreditation program requirements 2016. Available at: <http://www.acraccreditation.org/~media/ACRAccreditation/Documents/MRI/Requirements.pdf> [Accessed September 2016]
- [13] Fleiss JL. *Statistical Methods for Rates and Proportions*. New York: John Wiley and Sons; 1981.

Figure 1. Consensus on Diagnostic Findings: Chart depicting the percent of exams reporting the same interpretive findings. Aggregating all of the exams' reports together there were 49 distinct findings (pathology at a specific motion segment). None of these findings appeared in 100% of the reports and 32.7% of these findings only appeared once across all study exams' reports.

Figure 2. Disc Herniation Reported by Effect on Nerve Root or Thecal Sac: Depiction of how disc herniation was reported in each study exam across the patient's lumbar motion segments.

Figure 3. Example from the reference exam for grading central canal stenosis. (Left) At level of the L2 pedicles the area of the thecal sac measures approximately 241 sq. millimeters. (Right) At level of the L2-3 disc space the area of the thecal sac measures approximately 67 sq. millimeters. The reduction of the thecal sac is greater than 2/3 and graded as severe stenosis.

Table 1. Aggregated interpretive errors along with the reported variability of the radiologists' reports at the 10 study centers for each pathology. The table is sorted by increasing interpretive miss rate.

Area of Pathology	Reference Exam Findings	True Positives	False Positives	False Negatives	True Positive Rate (Sensitivity)	False Negative Rate (Miss Rate)
Anterior spondylolisthesis	1	9	0	1	90.0%	10.0%
Vertebral fracture	1	7	0	3	70.0%	30.0%
Neural foraminal stenosis	4	27	1	13	67.5%	32.5%
Facet degeneration	4	25	0	15	62.5%	37.5%
Disc degeneration	5	30	0	20	60.0%	40.0%
Central canal stenosis	2	11	8	9	55.0%	45.0%
Disc herniation	4	21	2	19	52.5%	47.5%
Nerve root involvement	4	11	3	29	27.5%	72.5%
Lateral recess stenosis	0	0	2	0	N/A	N/A

- 1 Table 2. Interpretive errors of each study exam report from the 10 study centers compared to the
 2 set of reference findings. The table is sorted by decreasing sensitivity.

Study Exam	Reference Exam Findings	True Positives	False Positives	False Negatives	Error Count (FP + FN)	True Positive Rate (Sensitivity)	False Negative Rate (Miss Rate)
Exam 7	25	18	2	7	9	72.0%	28.0%
Exam 10	25	17	1	8	9	68.0%	32.0%
Exam 8	25	16	1	9	10	64.0%	36.0%
Exam 4	25	16	3	9	12	64.0%	36.0%
Exam 3	25	15	0	10	10	60.0%	40.0%
Exam 5	25	15	2	10	12	60.0%	40.0%
Exam 9	25	14	1	11	12	56.0%	44.0%
Exam 1	25	11	2	14	16	44.0%	56.0%
Exam 2	25	10	1	15	16	40.0%	60.0%
Exam 6	25	9	3	16	19	36.0%	64.0%
Average \pm Std Dev		14.1 \pm 2.9	1.6 \pm 0.9	10.9 \pm 2.9	12.5 \pm 3.2	56.4% \pm 11.7	43.6% \pm 11.7

3