The Spine Journal 17 (2017) 554–561

Clinical Study

Variability in diagnostic error rates of 10 MRI centers performing lumbar spine MRI examinations on the same patient within a 3-week period

Richard Herzog, MD, FACR[a,b,*], Daniel R. Elgort, PhD[b], Adam E. Flanders, MD[c], Peter J. Moley, MD[a]

[a]*Hospital for Special Surgery, 535 E. 70th St, New York, NY 10021, USA*
[b]*Spreemo Quality Research Institute, 88 Pine St 11th Floor, New York, NY 10005, USA*
[c]*Thomas Jefferson University Hospital, 132 South 10th St, Suite 1080B, Main Building, Philadelphia, PA 19107, USA*

Received 12 July 2016; revised 22 September 2016; accepted 14 November 2016

Abstract

BACKGROUND CONTEXT: In today's health-care 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 radiologists' reports from multiple imaging centers performing a lumbar MRI examination 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: This is a prospective observational study comparing the interpretive findings reported for one patient scanned at 10 different MRI centers over a period of 3 weeks to each other and to reference MRI examinations performed immediately preceding and following the 10 MRI examinations.

PATIENT SAMPLE: The sample is a 63-year-old woman with a history of low back pain and right L5 radicular symptoms.

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).

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

FDA device/drug status: Not applicable.

Author disclosures: ***RH:*** Consulting: Spreemo (D), outside the submitted work. ***DRE:*** Nothing to disclose. ***AEF:*** Nothing to disclose. ***PJM:*** Nothing to disclose.

The disclosure key can be found on the Table of Contents and at www.TheSpineJournalOnline.com.

The study was approved by the Hospital for Special Surgery Institutional Review Board.

The study was funded by the Spreemo Quality Research Institute.

* Corresponding author. MRI Division, Department of Radiology and Imaging, Hospital for Special Surgery, 535 E. 70th St, New York, NY 10021, USA. Tel.: 929-575-5935.

E-mail address: herzogr@hss.edu (R. Herzog)

http://dx.doi.org/10.1016/j.spinee.2016.11.009

RESULTS: Across all 10 study examinations, 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 examinations and only one finding was reported in nine out of 10 study examinations. Of the interpretive findings, 32.7% appeared only once across all 10 of the study examinations' 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 examinations was 12.5±3.2 (both false-positives and false-negatives). The average false-negative count per examination was 10.9±2.9 out of 25 and the average false-positive count was 1.6±0.9, which correspond to an average true-positive rate (sensitivity) of 56.4%±11.7 and miss rate of 43.6%±11.7.

CONCLUSIONS: This study found marked variability in the reported interpretive findings and a high prevalence of interpretive errors in radiologists' reports of an MRI examination 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 his or her MRI examination and which radiologist interprets the examination may have a direct impact on radiological diagnosis, subsequent choice of treatment, and clinical outcome.

Keywords: MRI diagnostic accuracy; MRI diagnostic error rates; MRI diagnostic variability; MRI impact on patient outcome; MRI interpretive accuracy; MRI interpretive error rates; MRI interpretive variability; MRI lumbar spine; MRI quality; MRI value

Introduction

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

EVIDENCE & METHODS

Context

Variability in the readings of lumbar MRI scans by different radiologists using different MRI scanners is well appreciated by most of us involved in spine care. The authors set out to quantify this observation.

Contribution

They found poor agreement among readers and significant error rates.

Implications

While the choice of gold standard to assess error rates in such studies is often difficult, the lead author in this particular paper is certainly considered a leading expert in the field. Beyond that, the poor agreement among reviewers is clear, as are the implications. Treatments, insurance coverage, work status, etc., are all impacted by reported findings on MRI and variability is greatly concerning. Efforts to improve the situation are warranted.

Materials and methods

The study subject was a 63-year-old woman with a history of low back pain and right L5 radicular symptoms. Her pain radiated down to the anterolateral side of her right leg. On examination, the patient had mild weakness in right ankle dorsiflexion 4+/5 and right great toe extension 4+/5, reflexes were diminished, but symmetrical bilaterally (1/4), and she had a positive dural tension (seated slump) sign on the right. After an institutional review board approval, the subject underwent 12 MRI examinations of the lumbar spine. Ten examinations were performed at 10 different regional imaging centers over a period of 3 weeks, along with two reference MRI examinations performed at one of the authors' institutions immediately preceding and following the 10 MRI examinations. The reference examinations were performed on a closed 1.5T MRI system and included the following sequences: spin echo T1, spin echo T2, and short tau inversion recovery sagittal sequences (all sequences were acquired with a slice thickness of 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–L3 disc space), the second from L3 to S1 (parallel to the L4–L5 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 was 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 10 study centers performed their routine MRI examinations.

The study centers were selected by their location within or in 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, 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 any of the authors' institutions. The prescription stated that the patient had back and leg pain. If requested verbally, or as part of an intake questionnaire, the subject provided the same history of back pain and leg pain to each MRI center. The subject did not reveal that she was participating in a clinical study.

Following completion of the 10 study MRI examinations, the MRI reports from these examinations were stripped of all information identifying the center, radiologist, and type of equipment used. The reports were then reviewed by one of the authors, a subspecialized spine radiologist, and all the reported findings (appearing in either the Body or Impression sections) were inserted into a single "Study Exam Data Sheet" for cross-examination comparisons. Interpretive variability was then quantified using percent agreement rates and Fleiss kappa statistic.

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

For the purpose of which reference findings to use for the evaluation of interpretive errors, the authors limited the findings to a subset of findings that were reported in the Spinal Patient Outcomes Research Trial (SPORT) [9,10]. Specifically, the reference findings were limited to disc degeneration, disc herniation, spinal stenosis, nerve root involvement, facet degeneration, anterior spondylolisthesis, and vertebral fracture. The diagnosis of a disc herniation was based on detecting a localized or focal displacement of disc material beyond the limits of the intervertebral disc space [5]. The type of disc herniation, that is, protrusion, extrusion, or sequestered fragment, was not captured as many study examination 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 with 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 with the study examination findings collected in the Study Exam Data Sheet to identify interpretive errors. An error in interpretation in a study examination was considered present if there was no mention in the report of a reference finding. Any positive finding reported in a study examination that was not present in the reference findings was also considered an error in interpretation. There were no instances in which a positive finding reported on a study center examination was missed by the two independent reviewers during the evaluation of the reference examinations.

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 (eg, severe was present and only called mild). Similarly, to reduce overreporting 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 the reported findings across the 10 study examinations. Across all 10 examinations, there were 49 distinct findings reported (in either the Body or the

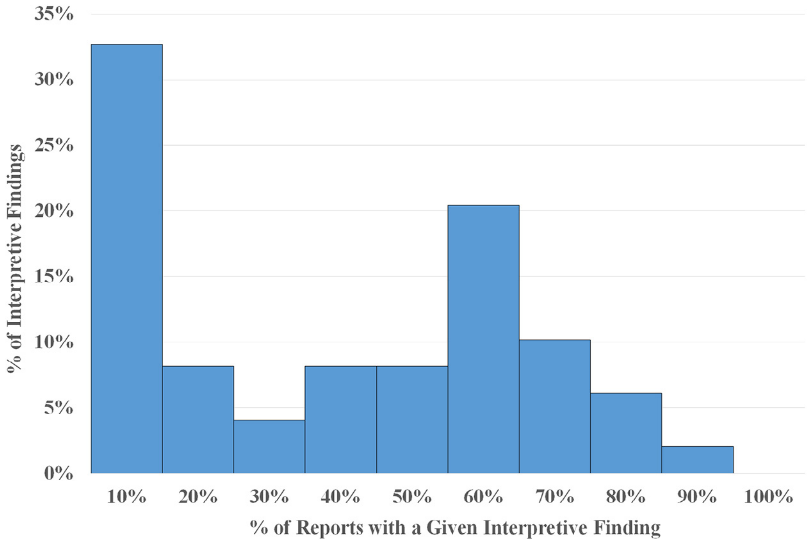

Fig. 1. Consensus on diagnostic findings: chart depicting the percent of examinations reporting the same interpretive findings. Aggregating all of the examinations' reports together, there were 49 distinct findings (pathology at a specific motion segment). None of these findings appeared in 100% of the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

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 and desiccation or spondylosis, disc height, disc herniation, stenosis of the central canal, lateral recess and neural foramina, nerve root involvement, end plate 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 examinations, and only one of the findings, the anterior spondylolisthesis present at L5–S1, was reported in 9 out of 10 examinations. Of the interpretive findings, 32.7% only appeared once across all 10 reports (Fig. 1).

The overall level of agreement on the reported findings of the study examinations 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 examinations' reports. A Fleiss kappa statistic value of 0 or less than 0 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 and 0.75 are interpreted as intermediate or good agreement, and values below 0.4 are interpreted as poor agreement [13]. The overall Fleiss kappa statistic across the 10 examinations and all reported interpretive findings was 0.20±0.03, indicating poor overall agreement on interpretive findings.

To illustrate the variation in the study examinations' reported interpretive findings, Fig. 2 depicts how a disc herniation was reported in the 10 examinations. The number of examinations reporting the presence of a disc herniation at a given motion segment ranged from 70% at L3–L4 to 20% at L5–S1; two examinations reported a disc herniation at all five motion segments and one examination did not report a disc herniation at any motion segment. The number of study examinations reporting thecal sac compression due to a disc herniation ranged from 60% at L1–L2 to only one examination reporting thecal sac compression at L4–L5. Nerve root involvement due to a disc herniation was reported in 20% of the examinations at L2–L3, 40% of the examinations at L3–L4, and 30% of the examinations at L4–L5. 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 examinations. The number of study examinations reporting the presence of central canal stenosis at a given motion segment ranged from 80% at L3–L4 to only one examination reporting central canal stenosis at L1–L2. Central canal stenosis was reported at four motion segments in two examinations and not present at any motion segment in two examinations. 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 5 out of the 10 examination reports included descriptions of any effect of spinal pathology on nerve roots. The number of study examinations reporting the presence of nerve root involvement at a given motion segment ranged from 50% at L3–L4 to only one examination reporting nerve root involvement at L5–S1. In four study examinations, nerve root involvement was reported at three motion segments, and in five study examinations 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

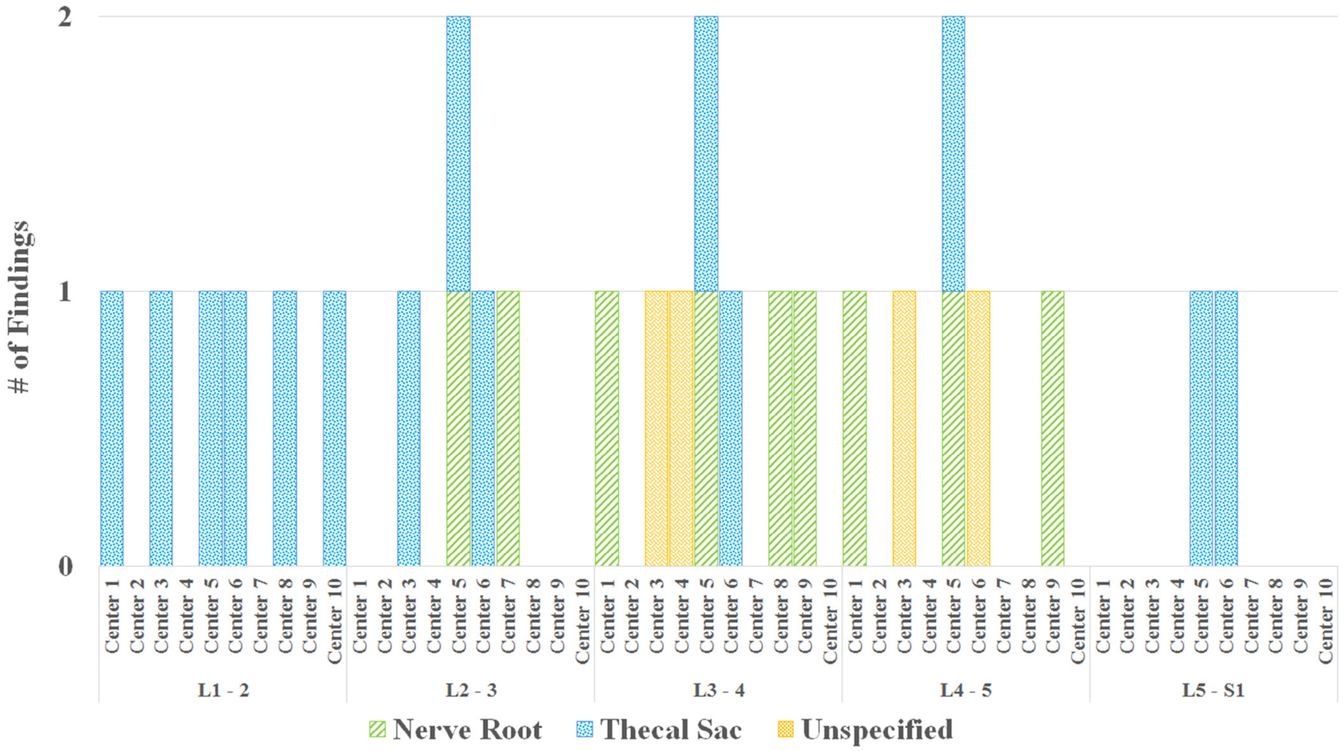

Fig. 2. Disc herniation reported by effect on nerve root or thecal sac: depiction of how disc herniation was reported in each study examination across the patient's lumbar motion segments.

study examinations, based on comparison of the study examinations with the reference examinations. The study examinations 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% and are summarized in Table 1.

Fig. 3 illustrates an example from the reference examination for grading central canal stenosis. At the level of the L2 pedicles, the area of the thecal sac measures approximately 241 mm^2, and at the level of the L2–L3 disc space the area of the thecal sac measures approximately 67 mm^2. The reduction of the thecal sac is greater than two-thirds and is graded as severe stenosis. At L2–L3, the central canal stenosis was not reported in four, reported as moderate in five, and severe in one study MRI examination (no error of interpretation was assessed for reporting the stenosis as moderate because, as indicated in the Materials and 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 three, moderate in four, and severe in one of the study MRI examinations.

Table 2 summarizes the interpretive errors of each study examination report compared with the set of reference findings. The study examinations' average interpretive error count was 12.5±3.2 per examination (both false-positives and

Table 1
Aggregated interpretive errors along with the reported variability of the radiologists' reports at the 10 study centers for each pathology

Area of pathology	Reference examination 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

The table is sorted by increasing interpretive miss rate.

Table 2
Interpretive errors of each study examination report from the 10 study centers compared with the set of reference findings

Study examination	Reference examination 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±standard deviation		14.1±2.9	1.6±0.9	10.9±2.9	12.5±3.2	56.4±11.7	43.6%±11.7

The table is sorted by decreasing sensitivity.

false-negatives). The study examinations' 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%±11.7 and false-negative rate (miss rate) of 43.6%±11.7.

Discussion

This study is the first in which interpretive variability and error rates were assessed across 10 complete lumbar MRI examinations of the same patient, conducted at 10 unaffiliated imaging centers within a 3-week period, and interpreted by radiologists who were blinded to their participation in the study. This study identified marked variability in the reported interpretive findings and an alarmingly high number of interpretive errors in the lumbar MRI reports. With respect to variability, no interpretive findings were reported in all 10 study examinations and only one finding was reported in 9 out of 10 study examinations. Of the interpretive findings, 32.7% only appeared once across all 10 of the study examinations' reports. A global Fleiss kappa statistic, computed across all reported findings, was 0.20±0.06, indicating poor overall agreement on interpretive findings. The level of variability across the examinations in this study is higher than the variability reported in previous studies in the literature which assessed variability of interpretation of the same set of images for multiple patients and grading a restricted set of pathologies and employing a pre-agreed set of definitions in most studies [7–11]. Quantifying the prevalence and types of interpretive errors in the study examinations, there

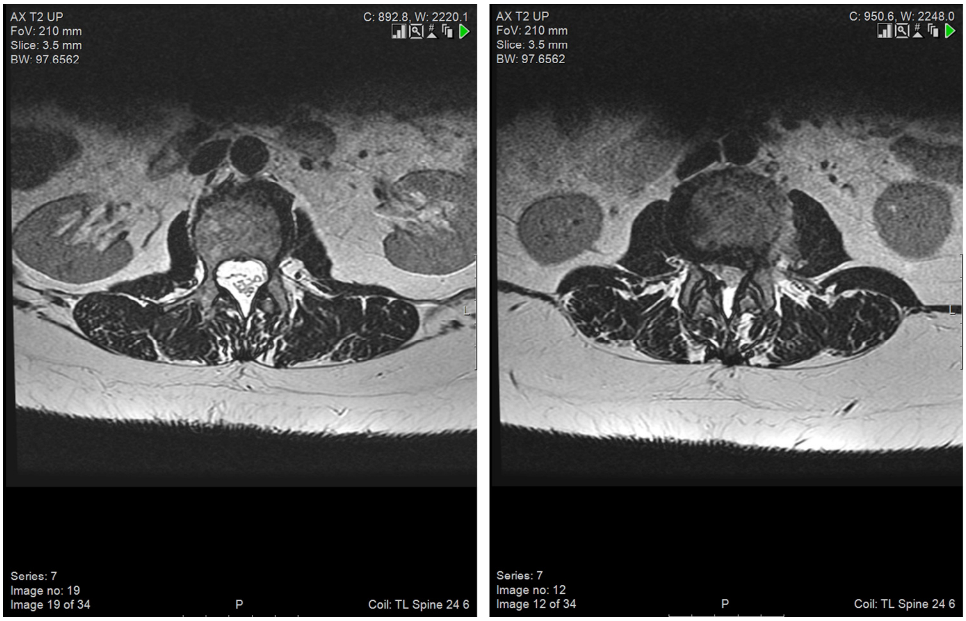

Fig. 3. Example from the reference examination for grading central canal stenosis. (Left) At the level of the L2 pedicles, the area of the thecal sac measures approximately 241 mm^2. (Right) At the level of the L2–L3 disc space, the area of the thecal sac measures approximately 67 mm^2. The reduction of the thecal sac is greater than two-thirds and was graded as severe stenosis.

was an average of 12.5±3.2 interpretive errors (both false-positives and false-negatives) across the 10 MRI examinations. The high average interpretive miss rate of 43.6%±11.7 across the study examinations means that important pathologies are routinely underreported. For example, the study examinations demonstrated an average miss rate for disc herniation of 47.5%. Similarly, the high false-positive rates for specific pathologies indicate that diagnostic findings, such as central canal stenosis, may be routinely overcalled.

The authors acknowledge that many physicians, in particular spine specialists, are generally able to independently review MRI examinations to verify the reported findings in order to formulate the most appropriate treatment plan including surgical care. But some physicians who provide care in the acute stages of a patient's symptoms (eg, family practice, internal medicine) are not as well-trained or experienced in reviewing MRI examinations. As a result, the initial diagnosis and treatment recommendations may be based on an inaccurate MRI interpretation, resulting in incorrect treatment recommendations, delayed recovery, or poor outcome. Moreover, for patients being considered for less invasive procedures (eg, interventional pain management or other minimally invasive procedures), there may be an overreliance on the MRI report. Importantly, it would be an omission not to add that the payer community heavily relies upon MRI reports during utilization and authorization review procedures. As a result, an incorrect diagnosis on an MRI has the potential to significantly delay authorization of appropriate care that in turn can negatively impact patient outcomes.

Several limitations of the current investigation exist due to study design and practical constraints. The first limitation is that because only a single MRI examination was performed at each of the 10 study MRI centers, the results reflect a single radiologist at each study center and may not be reflective of the overall performance of the MRI center. For this reason, the authors were unable to evaluate whether or not these findings were representative of the imaging centers selected for the study or generalizable to other MRI centers. Second, the sample size and geographic distribution of the study centers needed to be restricted in this study for logistical reasons to ensure the centers were accessible to the patient and could all be visited within a short time frame, as well as to respect the limits of the patient's tolerance for repeated examinations. Third, the authors selected a set of study centers employing a range of equipment types to reflect the variation present in the regional market of the subject. This distribution may deviate from the true distribution outside of the study area. Moreover, because many of the equipment types were only used for one or two of the examinations, there were not sufficient sample sizes to make statistically meaningful conclusions about the impact of MRI scanner type (eg, 0.3T vs. 1.5T vs. 3.0T) on the variability or interpretive error rates. Fourth, for similar reasons as above, the study was unable to evaluate the correlation between variability or errors and examination cost or other characteristics that may have varied across study centers and radiologists.

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

Because of these limitations, the authors did not attempt to identify or assess the relative importance of factors that explain the observed variability and errors across the 10 study examinations. However, potential reasons for the variability in the interpretation of the MRI examinations and prevalence of interpretive errors include the degree of specialization of the radiologist interpreting the MRI examination, the type of equipment and imaging sequences used at the study centers, and the nomenclature employed by the radiologists to describe and communicate abnormalities detected on the MRI examination. The authors did not attempt to train the radiologists at the 10 study centers on spinal nomenclature, as the study was designed to simulate what is currently occurring in the medical community where there is little agreement on the nomenclature used to describe many spinal pathologies. Moreover, the omission or inclusion of pathologic findings may vary based on community standards for a variety of reasons, including but not limited to the opinions of the referring physicians and the interpreting radiologists as to how distinct findings may be contributory to a patient's symptoms. The authors acknowledge that the potential reasons cited for the variability in the interpretation are speculative, and additional important factors may also be contributing to the observed variability.

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

Based on the variability and interpretive errors identified in this study, further investigation is required to understand the causes of these findings and their impact on the trajectory of patient care, outcomes, and costs. Moreover, awareness of the prevalence of errors may benefit providers in circumstances where there is poor correlation between a patient's clinical presentation and the reported MRI findings. Ultimately, it is the authors' opinions that accurate and complete diagnostic information at the onset of an injury or illness is critical to improve the chances for a patient's full recovery. However, reducing diagnostic errors and variability in reported findings will require the development and adoption of systematic mechanisms for measuring diagnostic MRI quality, including error rates. The authors acknowledge that accurately measuring interpretive errors at scale is a significant challenge and that some health-care providers may be reluctant to adopt such a system due to concerns around exposure of their errors, negative impact on reimbursement, and potential liability. Broad acceptance of the prevalence of errors and their potential impact on care is a critical first step toward a system capable of providing industry-wide, standardized measurement of diagnostic MRI quality.

Acknowledgments

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

References

[1] Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. AJR Am J Roentgenol 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 Radiat Oncol 2012;56:173–8. 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;8:934–8. 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 Orthop Relat Res 2015;473:1896–902. doi:10.1007/s11999-014-3674-y.

[7] Fu MC, Buerba RA, Long WD, Blizzard DJ, Lischuk AW, Haims AH, et al. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. Spine J 2014;14:2442–8. 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. Global Spine J 2015;5:406–10. http://dx.doi.org/10.1055/s-0035-1551651.

[9] Lurie JD, Tosteson AN, Tosteson TD, Carragee E, Carrino J, Kaiser J, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine 2008;33:1605–10. doi:10.1097/BRS.0b013e3181791af3.

[10] Carrino JA, Lurie JD, Tosteson ANA, Tosteson TD, Carragee EJ, Kaiser J, et al. Lumbar spine: reliability of MR imaging findings. Radiology 2009;250:161–70.

[11] Speciale AC, Pietrobon R, Urban CW, Richardson WJ, Helms CA, Major N, et al. 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.