Reliability Analysis of Deep Learning Algorithms for Reporting of Routine Lumbar MRI Scans

Kai-Uwe Lewandrowski, Narendran Muraleedharan, Steven Allen Eddy, Vikram Sobti, Brian D. Reece, Jorge Felipe Ramírez León and Sandeep Shah

Int J Spine Surg 2020, 14 (s3) S98-S107
doi: https://doi.org/10.14444/7132
https://www.ijssurgery.com/content/14/s3/S98

The International Journal of Spine Surgery
2397 Waterbury Circle, Suite 1,
Aurora, IL 60504, Phone: +1-630-375-1432

International Journal of Spine Surgery, Vol. 14, Supplement 3, 2020, pp. S98–S107
https://doi.org/10.14444/7132

Reliability Analysis of Deep Learning Algorithms for Reporting of Routine Lumbar MRI Scans

KAI-UWE LEWANDROWSKI, MD,[1] NARENDRAN MURALEEDHARAN, BASME,[2] STEVEN ALLEN EDDY, MD,[3] VIKRAM SOBTI, MD, MBA,[4] BRIAN D. REECE, MD,[5] JORGE FELIPE RAMÍREZ LEÓN, MD,[6] SANDEEP SHAH, MSEE, MBA[7]

[1]Staff Orthopaedic Spine Surgeon Center for Advanced Spine Care of Southern Arizona and Surgical Institute of Tucson, Tucson, Arizona, [2]Aptus Engineering, Inc, Scottsdale, Arizona, and Multus Medical, LLC, Phoenix, Arizona, [3]Multus Medical, LLC, Phoenix, Arizona, [4]Innovative Radiology, PC, River Forest, Illinois, [5]The Spine and Orthopedic Academic Research Institute, Lewisville, Texas, [6]Fundación Universitaria Sanitas, Bogotá, Colombia, Research Team, Centro de Columna. Bogotá, Colombia, Centro de Cirugía de Mínima Invasión, CECIMIN—Clínica Reina Sofía, Bogotá, Colombia, [7]Multus Medical, LLC, Phoenix, Arizona

ABSTRACT

Background: Artificial intelligence could provide more accurate magnetic resonance imaging (MRI) predictors of successful clinical outcomes in targeted spine care.

Objective: To analyze the level of agreement between lumbar MRI reports created by a deep learning neural network (RadBot) and the radiologists' MRI reading.

Methods: The compressive pathology definitions were extracted from the radiologist lumbar MRI reports from 65 patients with a total of 383 levels for the central canal: (0) no disc bulge/protrusion/canal stenosis, (1) disc bulge without canal stenosis, (2) disc bulge resulting in canal stenosis, and (3) disc herniation/protrusion/extrusion resulting in canal stenosis. For both, neural foramina were assessed with either (0) neural foraminal stenosis absent or (1) neural foramina stenosis present. Reporting criteria for the pathologies at each disc level and, when available, the grading of severity were extracted, and the Natural Language Processing model was used to generate a verbal and written report. The RadBot report was analyzed similarly as the MRI report by the radiologist. MRI reports were investigated by dichotomizing the data into 2 categories: normal and stenosis. The quality of the RadBot test was assessed by determining its sensitivity, specificity, and positive and negative predictive value as well as its reliability with the calculation of the Cronbach alpha and Cohen kappa using the radiologist MRI report as a gold standard.

Results: The authors found a RadBot sensitivity of 73.3%, a specificity of 88.4%, a positive predictive value of 80.3%, and a negative predictive value of 83.7%. The reliability analysis revealed the Cronbach alpha as 0.772. The highest individual values of the Cronbach alpha were 0.629 and 0.681 when compared to the MRI report by the radiologist, rending values of 0.566 and 0.688, respectively. Analysis of interobserver reliability rendered an overall kappa for the RadBot of 0.627. Analysis of receiver operating characteristics (ROC) showed a value of 0.808 for the area under the ROC curve.

Conclusions: Deep learning algorithms, when used for routine reporting in lumbar spine MRI, showed excellent quality as a diagnostic test that can distinguish the presence of neural element compression (stenosis) at a statistically significant level ($P < .0001$) from a random event distribution. This research should be extended to validated and directly visualized pain generators to improve the accuracy and prognostic value of the routine lumbar MRI scan for favorable clinical outcomes with intervention and surgery.

Level of Evidence: 3.

Clinical Relevance: Validity, clinical teaching, and evaluation study.

Special Issue

Keywords: artificial intelligence, deep neural network learning, magnetic resonance imaging, spinal pathologies, reliability analysis

INTRODUCTION

Minimally invasive and endoscopic transforaminal decompression techniques have become popular in spinal surgery due to technological advances.[1–6] There has been a substantial increase in the number of these types of procedures being carried out in ambulatory surgery centers.[7] The advantages of endoscopic transforaminal decompression are fewer postoperative complications, a shorter interval for return to work and social reintegration,[2,8–11] faster postoperative narcotic independence, and an overall reduced utilization of painkillers.[2,12] The latter problem is of significance in light of the opiate

abuse epidemic in the United States,[13] more rigorous medical necessity assessment,[14] and a demand for value-based health care measures to serve the aging baby-boomer population.[13–15] In this context, a conclusive preoperative diagnostic work-up of lumbar radiculopathy is crucial, as decompression is often limited to a small area of 1 affected neuroforamen and lateral recess.[16–18]

In this article, the authors report on the feasibility of using a deep learning algorithm for routine reporting in spine magnetic resonance imaging (MRI). The ultimate objective of this research is to improve the accuracy and predictive value of the MRI scan when applied to the preoperative planning of targeted minimally invasive and endoscopic spinal surgeries. These targeted procedures often ignore the majority of pathologies reported on routine lumbar MRI scans of patients with injuries or degenerative conditions of the spine and focus treatment only on the validated painful pathologies. The preoperative MRI scan is an integral part of the diagnostic work-up besides history, physical examination, electrodiagnostic studies, and confirmative diagnostic spinal injections.[15–19] The need to improve the diagnostic accuracy of the routine MRI scan has been well recognized by surgeons who reported on the correlation between intraoperatively observed findings as gold standard references and reflected on the use of the MRI scan as a predictor of the need for appropriate treatment and its clinical outcomes.[20–24] The MRI scan, in many respects, has become the ultimate gatekeeping test in the medical necessity determination of many spinal surgeries. Diagnostic inaccuracies related to false-negative diagnoses, therefore, have a significant impact on patient care and often lead to overutilization in other subspecialties of spine care, such as pain management. From a cost-benefit point of view, these inappropriate points-of-care interactions often translate into wasted treatments if considered ineffective by patients who continue to look for care but should be treated definitively by addressing the structural problems associated with their primary spinal pain generator. Therefore, improving the value of the MRI scan as a predictor of clinical outcomes with appropriate surgical treatments is not only central but also critical to applying the value-based approach to spine care. In this study, the authors report on the results of the sensitivity, specificity, and positive and negative predictive value; Cronbach alpha reliability; and interobserver Cohen kappa analysis of MRI reports produced by deep learning neural network algorithms when compared to routine reporting provided by the radiologist.

MATERIALS AND METHODS

The premise of this research and development is based on the ability for deep learning neural network models to identify features in MRI data that represent varying intensities or severities of degenerative pathologies or injuries in patients. The feasibility of this artificial intelligence (AI) approach was demonstrated in another study included in this journal's special focus issue. In this investigation, the same team of authors is now reporting on the statistics of the accuracy and reliability analysis with the AI approach to lumbar MRI reporting, which was considered the gold standard for the comparison analysis. All patients in this consecutive case series provided informed consent, and institutional review board approval was obtained (CEIFUS 106-19). Written informed consent was obtained from the patient for publication of this report and any accompanying images.

Patients and Training Data

The deep learning neural network models analyzed 65 lumbar MRI scans from the same number of patients, comprising a total of 383 levels. The DICOM data were ordered by the first author and were obtained from 1 MRI imaging center in patients with painful lumbar degenerative spine disease or injuries. The data set included the disc levels T12–L1, L1–L2, L2–L3, L3–L4, L4–L5, and L5–S1 for each patient. The average age of the 65 patients was 42.2 years with a standard deviation of 11.8 years. There were 51.5% male and 48.5% female patients. The MRI imaging centers provided radiology reports prepared and approved by board-certified radiologists. Each radiologist was required to present a reading for the presence or absence of annular bulging[25] (circumferential, paracentral, posterior), disc herniation[26] (extrusion, protrusion, sequestration, fragmentation), central canal stenosis[27–29] (compromise of the thecal sac with presence or absence of ventral epidural fat), and foraminal stenosis[30] (compromise of the left, right, or both neural foramina and nerve roots) for each intervertebral level.

Table 1. Level distribution of spinal disc spaces read by the radiologist and AI deep network learning.

	Level Distributions			
Disc Level	Frequency	Percent	Valid Percent	Cumulative Percent
T12–L1	63	16.4	16.4	16.4
L1–L2	64	16.7	16.7	33.1
L2–L3	64	16.7	16.7	49.8
L3–L4	64	16.7	16.7	66.5
L4–L5	64	16.7	16.7	83.2
L5–S1	64	16.7	16.7	100
Total	383	100.0	100.0	

Abbreviation: AI, artificial intelligence.

Extraction of MRI Data

For each disc location, the following classes were extracted from the radiologist report for the central canal: (0) no disc bulge/protrusion/canal stenosis, (1) disc bulge without canal stenosis, (2) disc bulge resulting in canal stenosis, and (3) disc herniation/protrusion/extrusion resulting in canal stenosis. One of the following classes was also extracted for each of the left and right neural foramina: (0) neural foraminal stenosis absent or (1) neural foraminal stenosis present. An example is shown in Table 1, where at the L3–L4 location in the side-by-side comparison, the radiologist read was converted to class (3) for the central canal, (0) for the left neural foramina, and (1) for the right neural foraminal and matched by the algorithm model. For the purpose of the reliability analysis, these findings were dichotomized into 2 simple categories: normal and stenosis.

Statistical Analysis

For the clinical outcome analysis, descriptive statistics (mean and standard deviation), cross-tabulation statistics of sensitivity, specificity, positive and negative predictive value, and measures of association were computed for 2-way tables using IBM SPSS Statistics software (version 27.0). The Pearson χ^2 and the likelihood-ratio χ^2 tests were used as statistical measures of association. The Multus RadBot MRI sensitivity of accurately grading and detecting symptomatic nerve root compression (true positive rate) (TP) was calculated on the basis of the grading by the board-certified radiologist as the percentage of patients (MRI positives) among the stenosis patients who were correctly identified by the Multus RadBot as having symptomatic neural compression confirmed by a board-certified radiologist. False negatives (FN) were patients with neural compression identified by the radiologist whose Multus RadBot MRI grading was negative for stenosis (MRI negatives). Therefore, diagnostic Multus RadBot MRI sensitivity for predicting a successful clinical outcome from endoscopic transforaminal decompression procedure was calculated as follows:

$$\frac{\text{MRI positives by radiologist reporting (TP)}}{\text{TP} + \text{Multus RadBot MRI negative (FN)}}$$

The Multus RadBot MRI specificity (true negative [TN] rate) of accurately detecting the absence of symptomatic nerve root compression as demonstrated by the radiologist's MRI reading was calculated as the percentage of patients correctly identified as not having symptomatic neural compression. False positives (FP) were defined as Multus RadBot MRI positives without the radiologist having identified the neural compression. Therefore, diagnostic Multus RadBot MRI specificity of predicting a neural element compression was calculated as follows:

$$\frac{\text{MRI negatives without compression by radiologist (TN)}}{\text{TN} + \text{Multus RadBot MRI positives without radiologist reading compression (FP)}}$$

The positive and negative predictive values of the Multus RadBot reading of the lumbar MRI scan for agreeing with the reading of the board-certified radiologist with the presence or absence of compressive pathology (normal or stenosis) were calculated as follows:

$$\frac{\text{MRI positives with compressive pathology reported by the radiologist}}{\text{TP} + \text{FP Multus RadBot MRI positives without compressive pathology reported by radiologist}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Intraobserver reliability between the reading provided by the radiologist and the neural network deep learning algorithm (Multus RadBot) was done by Cronbach alpha computation and Cohen kappa analysis as a measure of agreement between the radiologist's grading of the lumbar MRI scan and the Multus RadBot's assessment of foraminal and central stenosis. The Cohen kappa was calculated from the observed and expected frequencies on the diagonal of a square contingency table. The overall quality of the Multus RadBot algorithm as a diagnostic test was assessed with the receiver

Table 2. Frequency distribution stenosis as read by the radiologist.

	Original MRI Report by Radiologist			
Finding	Frequency	Percent	Valid Percent	Cumulative Percent
Normal	150	39.2	39.3	39.3
Stenosis	232	60.6	60.7	100.0
Total	382	99.7	100.0	
Missing	1	.3		
Total	383	100.0		

Abbreviation: MRI, magnetic resonance imaging.

operating characteristics (ROC) with determination of the area under the curve employing the left-upper-corner method using a dichotomization protocol of classifying MRI scan readings per intervertebral disc level as either normal or stenotic.[31–34] The confidence intervals for the likelihood ratios were calculated using the "log method."[35,36]

RESULTS

The level frequency distribution observed in the 65 patients is summarized in Table 1. The radiologist detected the presence of neural element compressive pathology (stenosis) in 60.6% of scanned levels, whereas the Multus RadBot AI algorithm determined the presence of stenosis in 64.2%, of scanned levels (Tables 2 and 3). As listed in Table 4, the most common levels reported as stenotic by the radiologist were L2–L3 (79.7%), L3–L4 (79.7%), and L4–L5 (77.8%). The frequency distribution read out by the Multus RadBot (Table 5) was similar, with some variation at L2–L3 (59.4%), L3–L4 (87.5%), and L4–L5 (93.8%), suggesting that pathology at the L2–L3 level was underdiagnosed versus overdiagnosed at the L4–L5 level. These differences were statistically significant ($P < .0001$). The ROC analysis showed a value of 0.808 for the area under the ROC curve (AUC), indicating that the Multus RadBot is an excellent diagnostic test that can detect the presence of neural element compression (stenosis) at a statistically significant level ($P < .0001$) from a random event distribution (Figure).

Table 3. Frequency distribution stenosis as read by AI deep network learning.

	RadBot AI Deep Learning Network MRI Reading			
Finding	Frequency	Percent	Valid Percent	Cumulative Percent
Normal	137	35.8	35.8	35.8
Stenosis	246	64.2	64.2	100.0
Total	383	100.0	100.0	

Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

Table 4. Frequency distribution of normal versus stenosis diagnosis as read by the radiologist.

	MRI Report as Read by the Radiologist, n (% Reported Within the Level)		Total, n (% Reported Within the Level)
Disc Level	Normal	Stenosis	
T12–L1	36 (56.3)	28 (43.8)	64 (100.0)
L1–L2	24 (37.5)	40 (62.5)	64 (100.0)
L2–L3	13 (20.3)	51 (79.7)	64 (100.0)
L3–L4	13 (20.3)	51 (79.7)	64 (100.0)
L4–L5	14 (22.2)	49 (77.8)	63 (100.0)
L5–S1	50 (79.4)	13 (20.6)	63 (100.0)
Total	150 (39.3)	232 (60.7)	382 (100.0)

Abbreviation: MRI, magnetic resonance imaging.

The cross tabulation between the Multus RadBot and radiologist's readings of the lumbar MRI scan using the radiologist's report as a gold standard revealed a Multus RadBot sensitivity of 73.3%, a specificity of 88.4% (Table 6), a positive predictive value of 80.3%, and a negative predictive value of 83.7% (Table 7), with all the differences in these 2 cross tabulations being statistically significant. The reliability analysis revealed the Cronbach alpha as 0.772. When cross tabulated by intervertebral disc level differences, in reliability by level were found (Table 8). Through a process of elimination, it was determined that Multus RadBot's performance was most reliable at the L2–L3 and L3–L4 levels with the highest individual values for the Cronbach alpha of 0.629 and 0.681 when compared to the MRI report by the radiologist, rending values of 0.566 and 0.688, respectively (Table 8). Kappa analysis of interobserver reliability rendered an overall kappa for the Multus RadBot of 0.627, suggesting that the Multus RadBot AI algorithm performed at a high reliability level (Table 9). Again, the diagnostic recognition of the Multus RadBot was the most reliable at the L2–L3 and L3–L4 levels on kappa analysis, showing kappa values of 0.738, and 0.606, respectively.

DISCUSSION

The results of this study highlighted a small "difference in opinion" in the interpretation of

Table 5. Chi-square tests for frequency distribution of normal versus stenosis diagnosis as read by the radiologist.

	Value	*df*	Asymptotic Significance (2-Sided)
Pearson chi-square	77.257[a]	5	.000
Likelihood ratio	79.333	5	.000
N of valid cases	382		

[a]0 cells (0.0%) have expected count less than 5. The minimum expected count is 24.74.

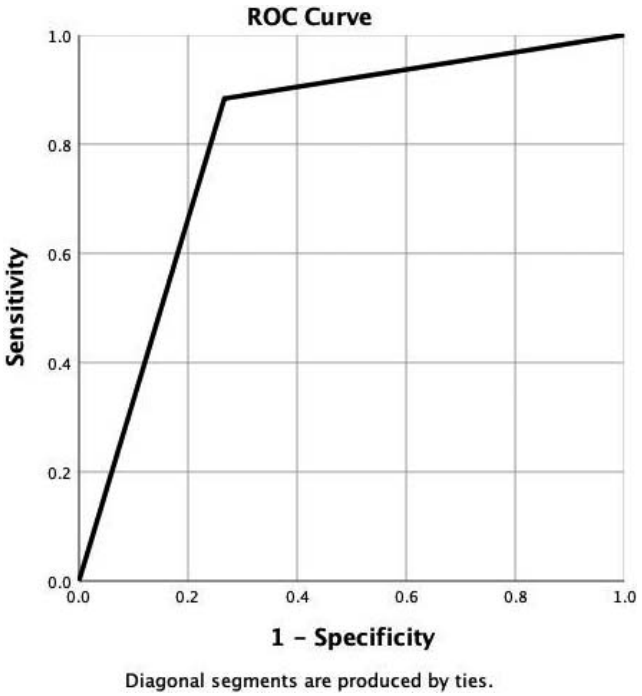

Figure. Area under the curve data = 0.808; SE = 0.025; asymptotic significance < 0.0001; asymptotic 95% confidence interval, lower bound = 0.760; upper bound = 0.857. Coordinates of the curve for artificial intelligence sensitivity = 0.884; 1 – specificity = 0.267.

routine lumbar MRI scans between the radiologist's report and AI deep neural network learning algorithm. While it is unclear whether the observed discrepancies arose out of the AI or the radiologist's reporting, it is obvious to see how such reporting discrepancies may impact patient selection for targeted spinal procedures, such as the endoscopic transforaminal surgery. As the determination of medical necessity in injured patients and in patients with painful degenerative conditions of the spine today hinges frequently on the exact verbatim reading in the MRI report, revisiting the accuracy of the MRI scan is of high relevance to patients and their physicians alike. False-positive readings may subject the patient to unwanted or unneeded

Table 6. Frequency distribution of normal versus stenosis diagnosis as read by RadBot.

	RadBot AI Deep Learning Network Reading		
Disc Level	**Normal**	**Stenosis**	**Total**
T12–L1	59 (93.7)	4 (6.3)	63 (100.0)
L1–L2	40 (62.5)	24 (37.5)	64 (100.0)
L2–L3	26 (40.6)	38 (59.4)	64 (100.0)
L3–L4	8 (12.5)	56 (87.5)	64 (100.0)
L4–L5	4 (6.3)	60 (93.8)	64 (100.0)
L5–S1	0 (0.0)	64 (100.0)	64 (100.0)
Total	137 (35.8)	246 (64.2)	383 (100.0)

Abbreviation: AI, artificial intelligence.

Table 7. Chi-square tests for frequency distribution of normal versus stenosis diagnosis as read by RadBot.

	Value	***df***	**Asymptotic Significance (2-Sided)**
Pearson chi-square	187.425[a]	5	.000
Likelihood ratio	220.410	5	.000
N of valid cases	383		

[a]0 cells (0.0%) have expected count less than 5. The minimum expected count is 22.54.

treatments at high expense, and false-negative interpretations may deny justified care. The consequences of this diagnostic dilemma play out every day, affecting individualized spine care of those patients with an estimated 2.06 million episodes of low back injury per year in the United States.[37]

The authors purposely chose a simplified way of analyzing the level of agreement between our AI and the radiologist's MRI reading by applying the following assumptions: (1) the MRI report by the radiologist was employed as the gold standard in this reliability and accuracy analysis, and (2) the authors categorized the MRI findings in a straightforward 2-category manner (normal anatomy or stenosis present) to facilitate the study of the AI algorithm's performance as a diagnostic test by employing accepted statistical methods of chi-square testing to determine the sensitivity, specificity, positive and negative predictive value, the overall test reliability with the ROC and AUC method or the calculations of the Cronbach alpha and Cohen kappa. The numbers obtained with these methods suggest that the our AI deep learning network as a diagnostic tool has excellent performance characteristics. Typically, Cohen kappa values of 0.6 and alpha over 0.7 and ROC values higher than 0.8 are considered the hallmarks of a highly useful diagnostic test.[38,39] It is not entirely clear to the authors why our AI deep learning neural

Table 8. Sensitivity and specificity of RadBot AI read versus MRI read by radiologist.

	Radiologist MRI Read		
RadBot	**Normal**	**Stenosis**	**Total**
Normal			
Count	110	27	137
% sensitivity within MRI scan	73.3	11.6	35.9
Stenosis			
Count	40	205	245
% specificity within MRI scan	26.7	88.4	64.1
Total			
Count	150	232	382
% within MRI scan	100.0	100.0	100.0

Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

Table 9. Chi-square tests for sensitivity and specificity of RadBot AI read versus MRI read by radiologist.[a]

	Value	*df*	Asymptotic Significance (2-Sided)
Pearson Chi-square	150.751[a]	1	.000
Likelihood ratio	148.081	1	.000
N of valid cases	383		

[a]0 cells (0.0%) have expected count less than 5. The minimum expected count is 53.80.
Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

Table 11. Chi-square tests for positive and negative predictive value of RadBot AI read versus MRI read by radiologist.[a]

	Value	*df*	Asymptotic Significance (2-Sided)
Pearson chi-square	150.751[a]	1	.000
Likelihood ratio	148.081	1	.000
N of valid cases	383		

[a]0 cells (0.0%) have expected count less than 5. The minimum expected count is 53.80.
Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

network was most accurate at the L2–L3 and L3–L4 levels. The most reasonable explanation is that pathologies at the other levels, but particularly at the L4–L5 level, are much more common, thus contributing to more significant variability in how these pathologies are read by the radiologist or interpreted by the Multus RadBot.

The authors are entirely aware of the limitation of their simplified statistical analysis by assuming that the MRI report provided by the reading radiologist was flawless. The authors could have chosen to have the radiologist's report reread by another 1 or 2 radiologists to incorporate that in the reliability discussion. However, the authors purposely decided against it so as not to create an artificial scenario that does not exist in the "real world," where routine lumbar MRI scans are read by 1 board-certified radiologist with little additional scrutiny. Clinical decision making affecting individual patients' lives are made like that every day. Therefore, the authors did not want to deviate from their simple side-by-side, Multus RadBot versus radiologist analysis approach. It goes without saying, though, that MRI raters on all sides of the medical necessity equation may use different radiological classification systems during the preoperative and diagnostic decision algorithms.[15,29,40,41] The first

Table 10. Positive and negative predictive value of RadBot AI read versus MRI read by radiologist.

	Regular MRI		
RadBot	Normal	Stenosis	Total
Normal			
Count	110	27	137
% positive predictive value within MRI scan	80.3	19.7	100.0
Stenosis			
Count	40	205	245
% negative predictive value within MRI scan	16.3	83.7	100.0
Total			
Count	150	232	382
% within MRI scan	39.3	60.7	100.0

Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

author has demonstrated this clinical dilemma affecting hundreds of his patients who were classified by the radiologist as false negatives but ultimately underwent successful transforaminal endoscopic decompression with excellent and good Macnab outcomes in over 88.3% of patients.[23] In his study of 1839 patients, the first author found a diagnostic gap of approximately 18% (330 patients),[24] which initially led to the denial of appropriate spine care by the patients' medical insurance. However, patients who persevered eventually underwent seemingly inappropriate endoscopic surgical decompression for their sciatica, back, and leg pain with a 94.6% success rate.[23] This type of spine care, deemed as medically not necessary based on traditional image-based clinical decision criteria done in patients responsive to successful endoscopic decompression, stimulated the authors of this study to look further into improving the preoperative diagnostic process in patients with sciatica due to herniated disc or stenosis leading up to targeted surgical decompression. Interestingly, this 18% diagnostic gap is commensurate with the Multus RadBot's percentage gain in reporting consistency in terms of sensitivity, specificity, and positive predictive value of the lumbar MRI scan with intervention reported by clinical studies where numbers are in the 60%–70% range.[18]

While the authors are encouraged by the excellent diagnostic performance parameters of the Multus RadBot's self-learning deep neural network models, they are also keenly aware of the underlying limitation of their study because of the underlying reporting bias inherent to the MRI reporting provided by the radiologists. Affective (unconscious emotional reaction) and cognitive (distortions of thinking) biases in the clinical diagnostic decision-making process may have impacted the radiologist's choice of words when dictating the findings he saw on the individual axial and sagittal MRI scan images.[42] Cognitive biases, such as hindsight or

Table 12. Interitem correlation matrix for reliability statistics of RadBot AI read versus MRI read by radiologist. The Cronbach alpha based on standardized items = .772 for normal and stenosis.

	RadBot					MRI					
	T12–L1	**L1–L2**	**L2–L3**	**L3–L4**	**L4–L5**	**T12–L1**	**L1–L2**	**L2–L3**	**L3–L4**	**L4–L5**	**L5–S1**
RadBot											
T12–L1	1.000	.064	−.055	.105	.071	.484	.045	−.213	.141	−.022	−.169
L1–L2	.064	1.000	.154	.007	.211	.051	.377	−.040	−.168	−.001	.030
L2–L3	−.055	.154	1.000	.180	−.082	−.032	.276	.720	.231	−.099	.113
L3–L4	.105	.007	.180	1.000	−.105	.100	.033	.312	.627	−.087	−.100
L4–L5	.071	.211	−.082	−.105	1.000	.147	−.181	−.065	−.141	.508	.011
Open											
T11–L1	.484	.051	−.032	.100	.147	1.000	.093	−.153	.003	.003	−.161
L1–L2	.045	.377	.276	.033	−.181	.093	1.000	.082	−.048	−.130	−.013
L2–L3	−.213	−.040	.720	.312	−.065	−.153	.082	1.000	.271	−.064	.153
L3–L4	.141	−.168	.231	.627	−.141	.003	−.048	.271	1.000	−.178	−.099
L4–L5	−.022	−.001	−.099	−.087	.508	.003	−.130	−.064	−.178	1.000	.092
L5–S1	−.169	.030	.113	−.100	.011	−.161	−.013	.153	−.099	.092	1.000

Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

outcome bias, are virtually unavoidable in a retrospective reclassification of clinical parameters, as knowledge of the outcome by the stakeholders in the patient care equation has been recognized to inflate the predictability of an event after it happened.[43,44] Hindsight cognitive biases may also have impacted the extent of disagreement in preoperative lumbar MRI grading by the radiologist.[45] Intuition bias may have played a role in the radiologist's wording of the MRI report while loosely adhering to radiographic stenosis classification systems.[45] The Multus RadBot is not subject to these biases for which reasons the authors expect higher reliability numbers incorrectly identifying painful spinal pathology with further refinements of the technology when directly visualized intraoperative observations of painful spinal pathologies are used as a gold standard rather than a radiologist report of another imaging modality. The first author has successfully used this approach in a prior study of the positive predictive value of the routine lumbar MRI scan.

Table 13. Total reliability statistics of RadBot AI read versus MRI read by radiologist. The Cronbach alpha based on standardized items = .772 for normal and stenosis.

	Scale Mean if Item Deleted	**Scale Variance if Item Deleted**	**Corrected Item-Total Correlation**	**Squared Multiple Correlation**	**Cronbach Alpha if Item Deleted**
RadBot					
T12–L1	6.4000	2.990	.094	.300	.421
L1–L2	6.0833	2.620	.173	.308	.398
L2–L3	5.8667	2.219	.456	.606	.272
L3–L4	5.6000	2.685	.296	.474	.365
L4–L5	5.5333	2.999	.083	.384	.423
L5–S1	No statistics are computed because RadBot is a constant.				
MRI					
T12–L1	6.2333	2.860	.064	.293	.435
L1–L2	6.0500	2.591	.185	.272	.393
L2–L3	5.8333	2.412	.322	.625	.335
L3–L4	5.6833	2.762	.144	.483	.408
L4–L5	5.6833	3.034	−.050	.303	.471
L5–S1	5.7000	2.959	−.005	.080	.458

Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

CONCLUSIONS

This study set out to better understand how to utilize the lumbar MRI scan as a prognosticator of favorable clinical outcomes when selecting patients for targeted spine care, such as the endoscopic transforaminal decompression procedure, aiming to cure patients of the predominant pain generator causing pain and disability in the functional context at the time when the spine care is delivered. To employ the routine lumbar MRI scan as a more accurate prognosticator for successful spine care with high patient satisfaction, this AI deep learning neural network, in the authors' opinion, needs to be further refined by focusing the segmentation models on MRI image findings of intraoperatively verified and validated pain generators responsive to treat-

Table 14. Total statistics for highest reliability levels of RadBot AI read versus MRI read by radiologist. The Cronbach alpha based on standardized items = .772 for normal and stenosis.

	Scale Mean if Item Deleted	**Scale Variance if Item Deleted**	**Corrected Item-Total Correlation**	**Squared Multiple Correlation**	**Cronbach Alpha if Item Deleted**
RadBot					
L2–L3	2.2787	.904	.529	.537	.629
L3–L4	2.0000	1.200	.447	.424	.681
MRI					
L2–L3	2.2459	.855	.614	.561	.566
L3–L4	2.0820	1.110	.424	.407	.688

Abbreviations: AI, artificial intelligence; MRI, magnetic resonance imaging.

Table 15. Kappa analysis of interobserver reliability RadBot versus MRI-reading by intervertebral disc level.

RadBot Disc Level	MRI Radiologist Reading			Kappa	Approximate Significance
	Normal	Stenosis	Total		
Disc.T12-L1					
Normal					
Count	50	9	59	.414	< 0.001
Expected Count	46.8	12.2	59.0		
% within RadBot	84.7	15.3	100.0		
Stenosis					
Count	0	4	4		
Expected Count	3.2	.8	4.0		
% within RadBot	0.0	100.0	100.0		
Disc.L1-L2					
Normal					
Count	28	12	40	.355	.004
Expected Count	22.5	17.5	40.0		
% within RadBot	70.0	30.0	100.0		
Stenosis					
Count	8	16	24		
Expected Count	13.5	10.5	24.0		
% within RadBot	33.3	66.7	100.0		
Disc.L2-L3					
Normal					
Count	21	5	26	.738	< 0.001
Expected Count	9.8	16.3	26.0		
% within RadBot	80.8	19.2	100.0		
Stenosis					
Count	3	35	38		
Expected Count	14.3	23.8	38.0		
% within RadBot	7.9	92.1	100.0		
Disc.L3-L4					
Normal					
Count	7	1	8	.606	< 0.001
Expected Count	1.6	6.4	8.0		
% within RadBot	87.5	12.5	100.0		
Stenosis					
Count	6	50	56		
Expected Count	11.4	44.6	56.0		
% within RadBot	10.7	89.3	100.0		
Disc.L4-L5					
Normal					
Count	4	0	4	.415	< 0.001
Expected Count	.8	3.2	4.0		
% within RadBot	100.0	0.0	100.0		
Stenosis					
Count	9	51	60		
Expected Count	12.2	47.8	60.0		
% within RadBot	15.0	85.0	100.0		
Disc.L5-S1					
Stenosis					
Count	14	49	63	.000[a]	< 0.001
Expected Count	14.0	49.0	63.0		
% within RadBot	22.2	77.8	100.0		
Total					
Normal					
Count	110	27	137	.627	< 0.001
Expected Count	53.8	83.2	137.0		
% within RadBot	80.3	19.7	100.0		
Stenosis					
Count	40	205	245		
Expected Count	96.2	148.8	245.0		
% within RadBot	16.3	83.7	100.0		
Total					
Count	150	232	382		
Expected Count	150.0	232.0	382.0		
% within RadBot	39.3	60.7	100.0		

[a]No statistics are computed because RadBot is a constant.

ment. The authors are in the process of completing a pilot study on this very problem. Surgical translational research on intraoperatively visualized spinal pathology should focus on analyzing the effectiveness of MRI prognosticators with spine surgical interventions, such as endoscopy, using state-of-the-art measures of central, lateral recess, and neural foraminal stenosis on MRI to further determine how they impact the prognosis of surgical treatment for neurogenic claudication and lumbar radiculopathy.

REFERENCES

1. Kambin P, Gennarelli T, Hermantin F. Minimally invasive techniques in spinal surgery: current practice. *Neurosurg Focus*. 1998;4(2):e8.

2. Adogwa O, Parker SL, Bydon A, Cheng J, McGirt MJ. Comparative effectiveness of minimally invasive versus open transforaminal lumbar interbody fusion: 2-year assessment of narcotic use, return to work, disability, and quality of life. *J Spinal Disord Tech*. 2011;24(8):479–484.

3. Bini W, Miller LE, Block JE. Minimally invasive treatment of moderate lumbar spinal stenosis with the superion interspinous spacer. *Open Orthop J*. 2011;5:361–367.

4. Al-Khouja LT, Baron EM, Johnson JP, Kim TT, Drazin D. Cost-effectiveness analysis in minimally invasive spine surgery. *Neurosurg Focus*. 2014;36(6):E4.

5. Liu C, Zhou Y. Percutaneous endoscopic lumbar diskectomy and minimally invasive transforaminal lumbar interbody fusion for recurrent lumbar disk herniation. *World Neurosurg*. 2017;98:14–20.

6. Yuan C, Wang J, Zhou Y, Pan Y. Endoscopic lumbar discectomy and minimally invasive lumbar interbody fusion: a contrastive review. *Wideochir Inne Tech Maloinwazyjne*. 2018;13(4):429–434.

7. Lewandrowski KU. Incidence, management, and cost of complications after transforaminal endoscopic decompression surgery for lumbar foraminal and lateral recess stenosis: a value proposition for outpatient ambulatory surgery. *Int J Spine Surg*. 2019;13(1):53–67.

8. Tye EY, Anderson JT, Haas AR, et al. The timing of surgery affects return to work rates in patients with degenerative lumbar stenosis in a workers' compensation setting. *Clin Spine Surg*. 2017;30(10):E1444–E1449.

9. Wang X, Borgman B, Vertuani S, Nilsson J. A systematic literature review of time to return to work and narcotic use after lumbar spinal fusion using minimal invasive and open surgery techniques. *BMC Health Serv Res*. 2017;17(1):446.

10. Khan I, Bydon M, Archer KR, et al. Impact of occupational characteristics on return to work for employed patients after elective lumbar spine surgery. *Spine J*. 2019;19(12):1969–1976.

11. Lewandrowski KU, Ransom NA, Yeung A. Return to work and recovery time analysis after outpatient endoscopic lumbar transforaminal decompression surgery. *J Spine Surg*. 2020;6(Suppl 1):S100–S115.

12. Nicholson T, Maltenfort M, Getz C, Lazarus M, Williams G, Namdari S. Multimodal pain management

protocol versus patient controlled narcotic analgesia for postoperative pain control after shoulder arthroplasty. *Arch Bone Jt Surg*. 2018;6(3):196–202.

13. Drahos GL, Williams L. Addressing the emerging public health crisis of narcotic overdose. *Gen Dent*. 2017;65(5):7–9.

14. Guyer R, Musacchio M, Cammisa FP, Jr., Lorio MP. ISASS recommendations/coverage criteria for decompression with interlaminar stabilization - coverage indications, limitations, and/or medical necessity. *Int J Spine Surg*. 2016;10:41.

15. Milette PC. Classification, diagnostic imaging, and imaging characterization of a lumbar herniated disk. *Radiol Clin North Am*. 2000;38(6):1267–1292.

16. Geurts JW, Kallewaard JW, Richardson J, Groen GJ. Targeted methylprednisolone acetate/hyaluronidase/clonidine injection after diagnostic epiduroscopy for chronic sciatica: a prospective, 1-year follow-up study. *Reg Anesth Pain Med*. 2002;27(4):343–352.

17. Lee IS, Kim SH, Lee JW, et al. Comparison of the temporary diagnostic relief of transforaminal epidural steroid injection approaches: conventional versus posterolateral technique. *AJNR Am J Neuroradiol*. 2007;28(2):204–208.

18. Kreiner DS, Baisden J, Gilbert T, Shaffer WO, Summers JT. Re: Diagnostic tests the NASS stenosis guidelines. *Spine J*. 2014;14(1):201–202.

19. Lewandrowski KU. Successful outcome after outpatient transforaminal decompression for lumbar foraminal and lateral recess stenosis: the positive predictive value of diagnostic epidural steroid injection. *Clin Neurol Neurosurg*. 2018;173:38–45.

20. Ghosh S, Chaudhary V. Supervised methods for detection and segmentation of tissues in clinical lumbar MRI. *Comput Med Imaging Graph*. 2014;38(7):639–649.

21. Costa DN, Passoni NM, Leyendecker JR, et al. Diagnostic utility of a likert scale versus qualitative descriptors and length of capsular contact for determining extraprostatic tumor extension at multiparametric prostate MRI. *AJR Am J Roentgenol*. 2018;210(5):1066–1072.

22. Lee SH, Yun SJ, Jo HH, Kim DH, Song JG, Park YS. Diagnostic accuracy of low-dose versus ultra-low-dose CT for lumbar disc disease and facet joint osteoarthritis in patients with low back pain with MRI correlation. *Skeletal Radiol*. 2018;47(4):491–504.

23. Lewandrowski KU. Retrospective analysis of accuracy and positive predictive value of preoperative lumbar MRI grading after successful outcome following outpatient endoscopic decompression for lumbar foraminal and lateral recess stenosis. *Clin Neurol Neurosurg*. 2019;179:74–80.

24. Yeung AT, Lewandrowski KU. Retrospective analysis of accuracy and positive predictive value of preoperative lumbar MRI grading after successful outcome following outpatient endoscopic decompression for lumbar foraminal and lateral recess stenosis. *Clin Neurol Neurosurg*. 2019;181:52.

25. Stokes IA. Surface strain on human intervertebral discs. *J Orthop Res*. 1987;5(3):348–355.

26. Fenyo A, Shinis D, Shelef I, et al. [Lumbar disc herniation: protrusion, extrusion or bulge? The proper use of the terms - how and when will it be defined as a disease?]. *Harefuah*. 2019;158(12):807–811.

27. Yuan S, Zou Y, Li Y, Chen M, Yue Y. A clinically relevant MRI grading system for lumbar central canal stenosis. *Clin Imaging*. 2016;40(6):1140–1145.

28. Lee GY, Lee JW, Choi HS, Oh KJ, Kang HS. A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method. *Skeletal Radiol*. 2011;40(8):1033–1039.

29. Lee CK, Rauschning W, Glenn W. Lateral lumbar spinal canal stenosis: classification, pathologic anatomy and surgical decompression. *Spine (Phila Pa 1976)*. 1988;13(3):313–320.

30. Lee S, Lee JW, Yeom JS, et al. A practical MRI grading system for lumbar foraminal stenosis. *AJR Am J Roentgenol*. 2010;194(4):1095–1098.

31. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*. 1978;8(4):283–298.

32. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord*. 2006;7:82.

33. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ. Assessment of the minimum clinically important difference in pain, disability, and quality of life after anterior cervical discectomy and fusion: clinical article. *J Neurosurg Spine*. 2013;18(2):154–160.

34. Azimi P, Yazdanian T, Benzel EC. Determination of minimally clinically important differences for JOABPEQ measure after discectomy in patients with lumbar disc herniation. *J Spine Surg*. 2018;4(1):102–108.

35. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557–560.

36. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*. 2009;339:b2700.

37. Mehling WE, Gopisetty V, Bartmess E, et al. The prognosis of acute low back pain in primary care in the United States: a 2-year prospective cohort study. *Spine (Phila Pa 1976)*. 2012;37(8):678–684.

38. Roine J, Uusitalo L, Hielm-Bjorkman A. Validating and reliability testing the descriptive data and three different disease diagnoses of the internet-based DOGRISK questionnaire. *BMC Vet Res*. 2016;12:30.

39. Spade CM, Fitzsimmons K, Houser J. Reliability testing of the psychosocial vital signs assessment tool. *J Psychosoc Nurs Ment Health Serv*. 2015;53(11):39–45.

40. Lewandrowski KU. "Outside-in" technique, clinical results, and indications with transforaminal lumbar endoscopic surgery: a retrospective study on 220 patients on applied radiographic classification of foraminal spinal stenosis. *Int J Spine Surg*. 2014;8:26.

41. Lee S, Kim SK, Lee SH, et al. Percutaneous endoscopic lumbar discectomy for migrated disc herniation: classification of disc migration and surgical approaches. *Eur Spine J*. 2007;16(3):431–437.

42. Valat JP. Epidural corticosteroid injections for sciatica: placebo effect, injection effect or anti-inflammatory effect? *Nat Clin Pract Rheumatol*. 2006;2(10):518–519.

43. Chang MC, Lee DG. Outcome of transforaminal epidural steroid injection according to the severity of lumbar foraminal spinal stenosis. *Pain Physician*. 2018;21(1):67–72.

44. Zwaan L, Monteiro S, Sherbino J, Ilgen J, Howey B, Norman G. Is bias in the eye of the beholder? A vignette study

to assess recognition of cognitive biases in clinical case workups. *BMJ Qual Saf*. 2017;26(2):104–110.

45. Henriksen K, Kaplan H. Hindsight bias, outcome knowledge and adaptive learning. *Qual Saf Health Care*. 2003;12 Suppl 2:ii46-50.

Disclosures and COI: The views expressed in this article represent those of the authors and no other entity or organization. The first author has no direct (employment, stock ownership, grants, patents), or indirect conflicts of interest (honoraria, consultancies to sponsoring organizations, mutual fund ownership, paid expert testimony). He is not currently affiliated with or under any consulting agreement with any MRI vendor that the clinical research data conclusion could directly enrich. This manuscript is not meant for or intended to push any other agenda other than reporting the research data related on automated recognition of common painful spine pathologies by deep neural network learning. The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Corresponding Author: Kai-Uwe Lewandrowski, MD, Staff Orthopaedic Spine Surgeon Center for Advanced Spine Care of Southern Arizona and Surgical Institute of Tucson, AZ 85712. Phone: (520) 204-1495; Fax: (623) 218-1215; Email: busniess@tucsonspine.com.

Published 9 December 2020