# Feasibility of Deep Learning Algorithms for Reporting in Routine Spine Magnetic Resonance Imaging

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

This information is current as of May 2, 2025.

---

# Feasibility of Deep Learning Algorithms for Reporting in Routine Spine Magnetic Resonance Imaging

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 is gaining traction in automated medical imaging analysis. Development of more accurate magnetic resonance imaging (MRI) predictors of successful clinical outcomes is necessary to better define indications for surgery, improve clinical outcomes with targeted minimally invasive and endoscopic procedures, and realize cost savings by avoiding more invasive spine care.

**Objective:** To demonstrate the ability for deep learning neural network models to identify features in MRI DICOM datasets that represent varying intensities or severities of common spinal pathologies and injuries and to demonstrate the feasibility of generating automated verbal MRI reports comparable to those produced by reading radiologists.

**Methods:** A 3-dimensional (3D) anatomical model of the lumbar spine was fitted to each of the patient's MRIs by a team of technicians. MRI T1, T2, sagittal, axial, and transverse reconstruction image series were used to train segmentation models by the intersection of the 3D model through these image sequences. Class definitions 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. Both the left and right 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 a natural language processing model was used to generate a verbal and written report. These data were then used to train a set of very deep convolutional neural network models, optimizing for minimal binary cross-entropy for each classification.

**Results:** The initial prediction validation of the implemented deep learning algorithm was done on 20% of the dataset, which was not used for artificial intelligence training. Of the 17,800 total disc locations for which MRI images and radiology reports were available, 14,720 were used to train the model, and 3560 were used to validate against. The convergence of validation accuracy achieved with the deep learning algorithm for the foraminal stenosis detector was 81% (sensitivity = 72.4.4%, specificity = 83.1%) after 25 complete iterations through the entire training dataset (epoch). The accuracy was 86.2% (sensitivity = 91.1%, specificity = 82.5%) for the central stenosis detector and 85.2% (sensitivity = 81.8%, specificity = 87.4%) for the disc herniation detector.

**Conclusions:** Deep learning algorithms may be used for routine reporting in spine MRI. There was a minimal disparity among accuracy, sensitivity, and specificity, indicating that the data were not overfitted to the training set. We concluded that variability in the training data tends to reduce overfitting and overtraining as the deep neural network models learn to focus on the common pathologies. Future studies should demonstrate the accuracy of deep neural network models and the predictive value of favorable clinical outcomes with intervention and surgery.

**Level of Evidence:** 3.

**Clinical Relevance:** Feasibility, clinical teaching, and evaluation study.

## INTRODUCTION

In the last 5 years, the development of artificial intelligence (AI) has leapfrogged forward across several industries, ranging from quality control in various manufacturing areas to improved automation of production processes and enhanced diagnostics in medical applications.[1] Examples include voice and facial recognition,[2,3] landmark localization,[4] autonomic driving,[5–9] and a wide array of medical imaging modalities.[10–21] From a clinical standpoint, demonstrating the application of AI and its deep neural network learning algorithms is highly relevant and timely due to the ongoing debate on the necessity of advanced medical and surgical intervention while costs are rising.[22–25] Establishing new diagnostic imaging criteria with higher sensitivity, specificity, accuracy, and positive predictive value for favorable clinical outcomes with proposed interventions is considered by many to be key to the development of next-generation patient-centered and higher-quality health care service purchasing strategies in general and for the endoscopic spinal procedure in particular.[26–31] Using AI in closing the well-recognized diagnostic cap in routine lumbar magnetic resonance imaging (MRI) scanning is one such example.[32] Extracting higher-quality diagnostic value from the MRI scan is critical for successful endoscopic spinal surgery because it relies heavily on correctly identifying the pain generator responsible for the patient's symptoms.

The deployment of severity grading of traditional subjective visual analysis of advanced cross-sectional MRI imaging of the spine[33–35] by the radiologist not only leaves room for errors but has been shown to lead to the omission of appropriate spine care, which ultimately contributes to overuse in other areas.[36] Repetitive rounds of less-effective physical therapy, interventional, and medical care often provide only short-term pain relief and rarely definitively address the patients' disability because the pain generator stemming from the underlying structural abnormality of the spine has not been fixed.[37–43] As a result, the patients' disabilities continue and play out in their private and professional lives with decreased functional capacity due to poorly controlled pain, lack of strength, coordination, or insufficient endurance.[37–43] The cumulative societal burden due to missed work, ongoing use of medical services, and narcotic dependence[44,45] is on the radar of every stakeholder in government and the medical insurance and service industry.[46,47]

Out-of-control runaway costs will likely prompt more rationing of medical services in general and spine care in particular[48–51] unless clinical evidence is presented on how to realize cost savings with technology advancements[52–59] that produce more impactful, targeted, and durable solutions for patients that ultimately have the potential to flex down the spending curve during a time in which the demand for such spine care is expected to increase significantly with generations of aging baby boomers coming onto the Medicare rolls.[60,61] More accurate prognosticators of favorable clinical outcomes with spine intervention and surgery are critical to materializing such cost savings. The need for such cost savings will likely also translate into higher use of more targeted minimally invasive and endoscopic spinal decompression and reconstructive surgeries. To provide consistent clinical benefit, identifying the primary pain generator is of utmost importance, and applying routine MRI scanning with higher-level accuracy will probably become more relevant. Therefore, we investigated the feasibility of using deep learning algorithms for routine reporting in spine MRI with the ultimate objective of improving its accuracy and predictive value.

## MATERIALS AND METHODS

The premise of this research and development is based on the ability of deep learning neural network models to identify features in MRI data that represent varying intensities or severities of pathologies or injuries in patients.

### Patients and Training Data

The training dataset used to develop the neural network models includes lumbar MRI scans from 3560 patients, constituting a total of 17 800 levels. The training data were obtained from 168 different MRI imaging center locations around the United States. The dataset included the disc levels L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1 for each patient. The average age of the 3560 patients was 41.2 years, with a standard deviation of 14.9 years. There were 46% male and 51.9% female patients. The remaining 2.1% chose not to identify their gender. The participating 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
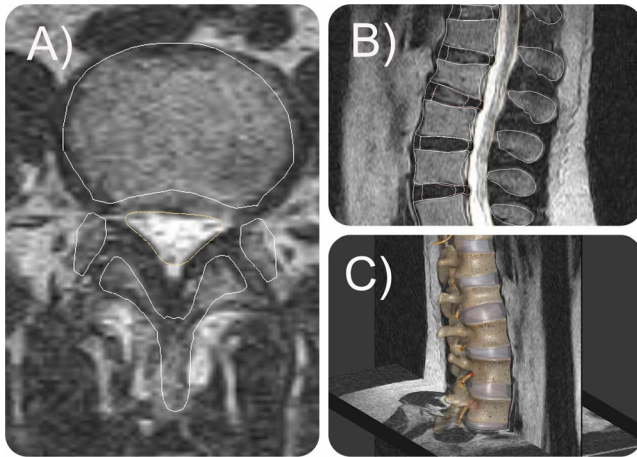
**Figure 1.** Example of segmentation of vertebrae, intervertebral discs, and dural sac on (a) sagittal and (b) axial MRI images and a corresponding (c) segmented 3-dimensional anatomical model. MRI, magnetic resonance imaging.

annular bulging[62] (circumferential, paracentral, posterior), disc herniation[63] (extrusion, protrusion, sequestration, fragmentation), central canal stenosis[33,64,65] (compromise of the thecal sac with presence or absence of ventral epidural fat), and foraminal stenosis[66] (compromise of the left, right, or both neural foramina and nerve roots) for each intervertebral level. For algorithm development and validation, various splits of the dataset were used to ensure that the model was tested against cases that it had not seen during training. Train-test percentages of 70%–30%, 80%–20%, and 90%–10% were used for various models.

## Preparation of Training Data

It is essential to extract numerical training data from the MRI imaging data. First, it is crucial to identify the location of each vertebra and disc in the patient's lumbar spine in order. Lu et al[67] proposed the use of automated segmentation algorithms to automate this process. In their approach, quadrilaterals were drawn to encompass each vertebra visible on the sagittal image. Segmented regions were used to fit a spine curve and localize the centers of each disc, and a series of sagittal and axial slices from the area was used for training and prediction.[67] To extract the disc regions more accurately and to extract the spinal cord profile, a three-dimensional (3D) anatomical model of the lumbar spine was fitted to each of the patient's MRIs by a team of technicians. The 3D model was fitted such that the boundaries of the vertebrae, discs, and cord line up with the respective boundaries in the MRI

images. Sagittal and axial slices were used as reference (Figure 1a and b). The segmentation results in a 3D anatomical model custom to each patient's lumbar spine (Figure 1c). This allows the use of other MRI image series (for example, T1, T2, sagittal, axial, and transverse reconstructions) to be used to train segmentation models by the intersection of the 3D model through these image sequences. For each disc location, the following classes were extracted from the radiologist report for the central canal: (0) no disc bulge, protrusion, or canal stenosis, (1) disc bulge without canal stenosis, (2) disc bulge resulting in canal stenosis, and (3) disc herniation, protrusion, or 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, (1) neural foramina stenosis present. An example is shown in Table 1, in which 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—which was matched by the algorithm model.

Second, 2 approaches were taken to extract manual radiologist reporting labels for the pathologies at each disc level and, when available, the grading of severity. Similar to Lu's DeepSPINE model83, natural language processing (NLP) was used to extract disc-level locations and pathologies at each location. The NLP model was trained with 5000 manually labeled disc levels. One of the following options was marked for the central canal on the basis of the radiologist's report: no signs of abnormality, disc bulging without compromise of the thecal sac, disc bulging compressing thecal sac (central canal stenosis), or disc herniation compressing thecal sac (central canal stenosis). One of the following options was labeled for the neural foramina as well: no signs of abnormality, left foraminal stenosis, right foraminal stenosis, or bilateral foraminal stenosis. For example, a report finding that states "L4-L5: Broad-based posterior disc herniation, best seen on sagittal T2 image #8/13 indenting thecal sac and causing mild narrowing of bilateral neural foramina" is labeled as follows: disc herniation compressing thecal sac (central canal stenosis) and bilateral foraminal stenosis.

The NLP algorithm was run on all 17 800 disc levels with radiology reports provided to generate labeled training data for the pathology identification

**Table 1.** Exemplary side-by-side comparison of the verbal report generated by the algorithm with radiologist's MRI report of the identical patient at disc levels L1-L2 through L5-S1.

| Level | Model Output | Radiologist Report |
|---|---|---|
| L1-L2 | No disc herniation, neurocompression, or neuroforaminal stenosis is seen at this level. | There is a 1-mm midline disc bulge. There is no central canal narrowing. |
| L2-L3 | There is moderate posterior disc herniation of the intervertebral disc impinging on the thecal sac. The spinal canal midline AP diameter is 11.8 mm. This is best seen on T2_FSE_TRS image slice No. 4. There is narrowing of the neural foramina bilaterally. | There is a moderate degree of central canal narrowing. There is a posterior annular tear. There is a 3-mm circumferential disc protrusion resulting in abutment of the descending L3 nerve roots bilaterally as well as abutment of the exiting right and left L2 nerve roots. |
| L3-L4 | There is posterior disc herniation of the intervertebral disc impinging on the thecal sac. The spinal canal midline AP diameter is 12.7 mm. This is best seen on T2_FSE_TRS image slice No. 9. There is narrowing of the right neural foramina with abutment of the exiting right nerve root. | There is a moderate degree of central canal narrowing. There is a 2-mm right foraminal disc protrusion with abutment of the exiting right L3 nerve root and mild narrowing of the right neural foramen. |
| L4-L5 | There is posterior annular bulging seen compressing the thecal sac, resulting in central canal stenosis. The spinal canal midline AP diameter is 15.9 mm. This is best seen on T2_FSE_TRS image slice No. 16. There is narrowing of the neural foramina bilaterally. | There is mild central canal narrowing. There is a 2-mm right foraminal disc protrusion resulting in mild abutment of the exiting right L4 nerve root with mild narrowing of the right neural foramen. |
| L5-S1 | There is posterior disc herniation of the intervertebral disc impinging on the thecal sac. The spinal canal midline AP diameter is 19.2 mm. This is best seen on T2_FSE_TRS image slice No. 23. There is narrowing of the neural foramina bilaterally. | There is a 1-mm midline disc bulge with mild effacement of the anterior thecal sac. There is no central canal narrowing. The height of the right neural foramen is 19 mm and the height of the left neural foramen is 17 mm. |

Abbreviations: AP, anterior-posterior.

deep learning algorithm. Due to known imperfections and accuracy of NLP algorithms, a semisupervised training process was adopted. Semisupervised training algorithms have been used to improve the accuracy of models when it is unfeasible to prepare supervised training data due to a large sample size or the complexity and labor intensiveness of manually labeling data.[68–70] The training process included unsupervised training data generated by the NLP algorithm for the entire dataset along with the 5000 manually labeled and curated labels prepared originally to train the NLP algorithm. Furthermore, due to the tendency for lumbar disc pathologies to be more common in the lower lumbar motion segments, class imbalance was handled by weighting model classes and mixing of disc locations in the training data. Figure 2 depicts
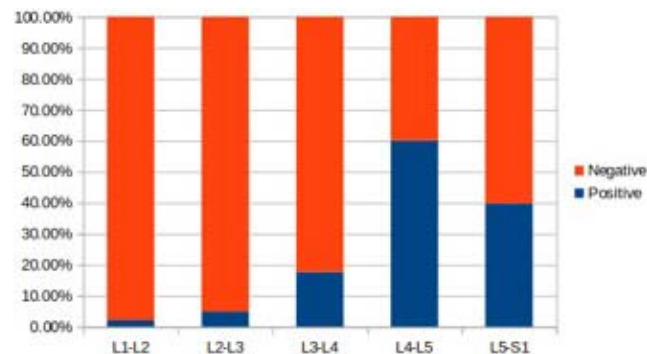


**Figure 2.** Positive and negative identifications of central canal stenosis at each disc level.

the distribution of identified central canal stenosis in the training data at each disc location. Suppressed consistency loss was used as a regularization method to increase robustness towards class imbalance at different levels.[71]

## Models and Architecture

The proposed algorithm operates in three high-level stages. First, each sagittal and axial slice is segmented using a semantic segmentation network using the manually segmented 3D model. The implemented method uses concepts from the one hundred layers tiramisu network proposed by Jégou et al.[72] Segmented outputs similar to those in Figure 3a and b are generated for each sagittal and axial slice in the MRI images. The segmented regions are used to extract the disc centers and orientation (using principal component analysis) for each disc location from L5-S1 counting upward until L1-L2. As proposed in the DeepSPINE model,[67] stacks of cropped sagittal and axial slices are extracted from MRI images intersecting the disc. The segmented spinal cord is also used to measure the canal midline anterior-posterior (AP) diameter—an objective and measurable metric. The second stage in the pipeline uses 2 separate visual geometry group convolutional networks[73] trained with semisupervised methods on cropped sagittal and axial MRI image stacks and radiological findings labeled using NLP and manually. The first network is used to detect and grade
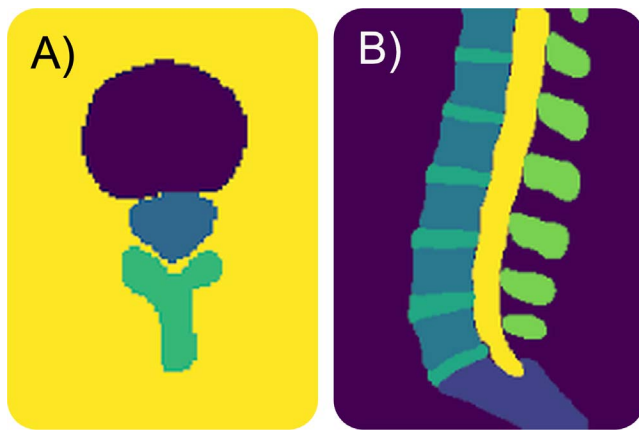
**Figure 3.** Representative (a) sagittal and (b) axial segmentation predictions of lumbar spine matching patient's MRI images in Figure 1. MRI, magnetic resonance imaging.

central canal stenosis, and the second to identify foraminal stenosis on the left and right neural foramen. The final stage compiles the predictions into a summary similar to that presented by radiologists and used to train the models. Simple decision trees are used to compile the summary. Differences in radiologist terminology and standards for detecting and grading stenosis affect the algorithm only minimally due to the same nomenclature and terminology used in the training data. A series of convolutional neural networks trained with gradient descent algorithms with dice loss coefficients and spatial dropout prevents overtraining to the dataset and enforces the network models (the RadBot) to identify defining features that result in diagnosis and grading. The same also enforces the network to ignore differences between radiologist terminology. Figure 4 depicts the high-level stages of the proposed algorithm, arriving at a printed MRI report generated by these deep learning algorithms.

## RESULTS

We compared the reports generated by these AI segmentation algorithms to select representative image sections and had the radiologist on our team review the images to confirm the AI read and to compare it with the radiologist report provided by the participating MRI centers. Figure 5a and b is an example diagnostic assessment using the algorithm that is known to have no disc bulging, no central canal stenosis, and no foraminal narrowing. The algorithm reported no canal stenosis and no neural foraminal stenosis, thus matching the known radiologist's reporting. The algorithm-generated report summary was "L1-L2: No disc herniation, neuro-compression, or neuroforaminal stenosis is seen at this level." Figure 5c and d is an example diagnostic assessment using the algorithm in which the training radiologist labeled the disc to have a posterior disc protrusion abutting the thecal sac and compromising the neural foramina bilaterally. In comparison, the deep learning algorithm reported "There is posterior herniation of the intervertebral disc impinging on the thecal sac, best seen on T2_FSE_TRS (FSE=fast spin echo) series image #4. The spinal canal midline AP diameter is 10 mm. There is narrowing of the neural foramina bilaterally." As demonstrated in these 2 examples, the algorithm also indicated the image slice in which the pathology was best demonstrated and reported the measured spinal canal diameter at the affected level.

We used the initial results of the prediction and validation of the implemented deep learning algorithm on the 20% of the dataset that was not used for the AI training. Of the 17,800 total disc locations for which MRI images and radiology reports were available, 14,720 were used to train the model, and
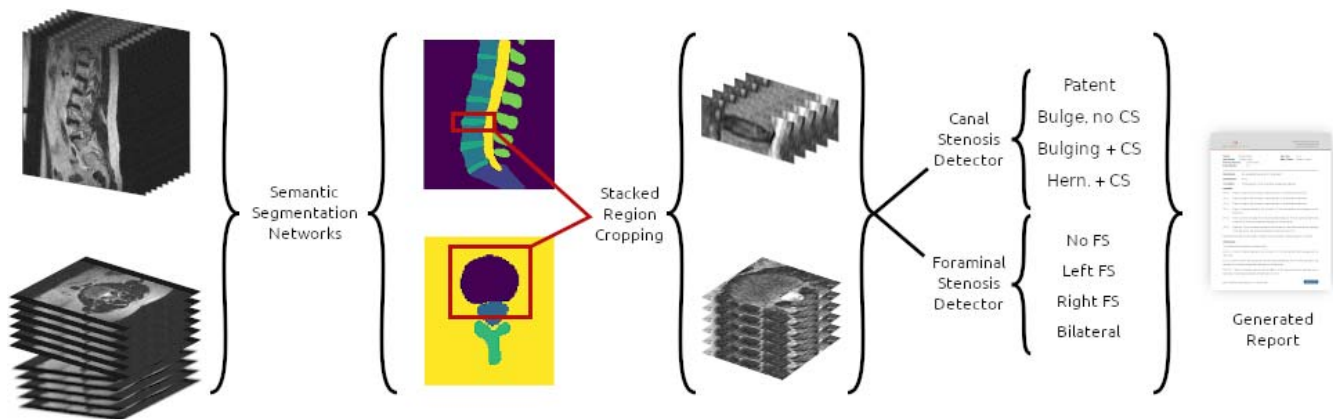


**Figure 4.** High-level architectural diagram of the implemented deep learning algorithm used to generate automated MRI reports. MRI, magnetic resonance imaging.
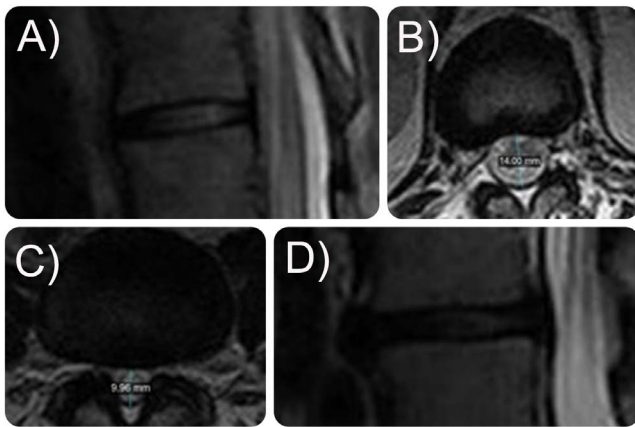
**Figure 5.** Exemplary L1-L2 MRI (a) sagittal and (b) axial images used for diagnostic assessment with the segmentation algorithm on levels known to have no disc bulging, no central canal stenosis, and no foraminal narrowing. Another example L1-L2 MRI (c) axial and (d) sagittal images of a diagnostic assessment using the segmentation algorithm in which the training radiologist labeled the disc to have a posterior disc protrusion abutting the thecal sac and compromising the neural foramina bilaterally.

3560 were used to validate against. Separate models were developed and trained for the identification of each diagnosis and class: generalized disc bulging, canal stenosis, disc herniation, and foraminal stenosis. The bilateral neuroforamina were assessed independently for stenosis affecting the left and right nerve roots. Therefore, twice as many data points were available to train the foraminal stenosis detector model. The loss functions used to minimize in each model was binary cross-entropy, also known as the log loss function.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \\ \cdot \log(1 - p(y_i)).$$

Each model was trained for 25 epochs, during which the convergence of binary validation accuracy was observed. Figure 6a is a plot depicting the convergence of validation accuracy achieved with the deep learning algorithm to approximately 81% for the foraminal stenosis detector. The optimization, however, was for the above-mentioned binary cross-entropy loss function. Figure 6b is a plot depicting the convergence of the binary cross-entropy loss across the 25 epochs (an *epoch* is 1 complete iteration through the entire training dataset) on the foraminal stenosis detector with increasing accuracy. At the end of each epoch, the binary cross-entropy loss was calculated and stochastic gradient descent optimization was used to compute changes to deep neural network model weights to minimize the loss. We observed from the plots that
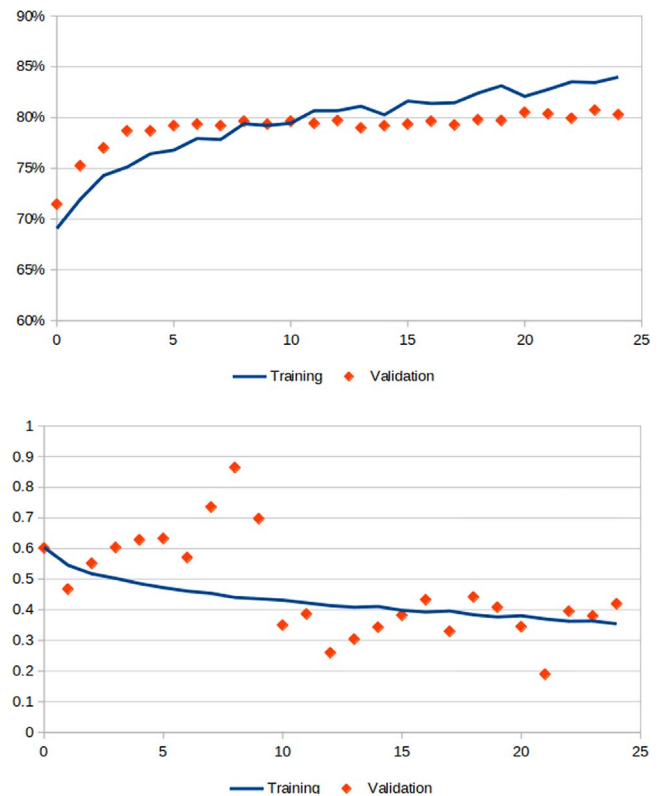


**Figure 6.** Graphic depiction of training convergence of the foraminal stenosis detector. In the top panel (6a), the x-axis is the number of training steps and the y-axis is the binary accuracy. The bottom panel (6b) shows a plot of binary cross-entropy (y-axis) versus number of training steps (x-axis). The binary cross-entropy is used to estimate error between the radiologist reads and the artificial intelligence predictions. Hence, decreasing binary cross-entropy is associated with desired accuracy gains.

the binary training accuracy continued to increase, whereas the validation accuracy converged to roughly 81% for the foraminal stenosis detector (sensitivity = 72.4.4%, specificity = 83.1%). The convergence of the validation accuracy was apparent after just 5 training epochs. Any gain in training accuracy observed past the validation accuracy convergence was due to overfitting to specific radiology reads and methods; however, these did not affect the overall validation accuracy. Spatial dropouts and other techniques were implemented to minimize overfitting to the specific training dataset. Binary accuracy, test sensitivity, and specificity were recorded for each model on the basis of the results from validating against 20% of the complete dataset; these are summarized in Table 2. The accuracy for the central stenosis detector was 86.2% (sensitivity = 91.1%, specificity = 82.5%) and for the disc herniation detector, 85.2% (sensitivity = 81.8%, specificity = 87.4%).

**Table 2.** Validation results showing accuracy, sensitivity, and specificity for each segmentation model.

| Detector Model | Accuracy, % | Sensitivity, % | Specificity, % |
|---|---|---|---|
| Bulging | 84.9 | 89.4 | 81.2 |
| Canal stenosis | 86.2 | 91.1 | 82.5 |
| Disc herniation | 85.2 | 81.8 | 87.4 |
| Foraminal stenosis | 81.0 | 72.4 | 83.1 |

## DISCUSSION

The need for more accurate prognosticators of favorable clinical outcomes with lumbar spinal surgery prompted us to investigate the feasibility of using deep learning algorithms for routine reporting in spine MRI. The need for accurate prognosticators of favorable clinical outcomes has been recognized by the North American Spine Society (NASS), which discussed this need in several of its consensus treatment guidelines for common spine problems.[74,75] The organization provided an in-depth review of the existing literature and graded the clinical evidence by comparing preoperative MRI findings with intraoperative observations directly visualized by the surgeon in open spine surgeries. Thus, vetted and validated sensitivity and specificity numbers ranging between the 60th and 70th percentiles and serving as industry benchmarks for most common spine problems were established.[74,75] One consequential example of the poor diagnostic value of routine lumbar MRI scans is missed injury to the posterior longitudinal ligament complex in thoracolumbar fractures in patients with acute injury. The integrity of this vital, stabilizing ligamentous complex typically triggers nonoperative care with bracing,[76] whereas injury triggers surgical fixation with spinal fusion[74,75]; an algorithm that calls for 2 vastly different treatments, when used erroneously, has tremendous unintended downstream consequences that nearly always translate into an ongoing need for care and higher cost. Another such example is herniated disc. A high percentage of asymptomatic, healthy volunteers were found to have disc herniations at multiple lumbar levels,[77] calling into question the positive predictive value of the lumbar MRI scan in patients with painful acute injuries or degenerative abnormalities.[36,78] In a nutshell, the lumbar MRI scan delivers little information with respect to the leading pain generator. The high false negative rate among patients with sciatica-type back and leg pain is on the order of 30%,[78] and yet the radiologists lacking

relevant clinical context of the spine care at the time it is delivered—willingly and knowingly or not—find themselves in the middle of the medical necessity controversy when it comes to determining the need for treatment.[43] What is evident is that there is a tremendous need to improve the accuracy of the interpretation of the MRI scan, particularly when it comes to the application of small, targeted, minimally invasive and endoscopic surgeries that aim to treat only the most relevant pain generator.[38,79,80] Higher preoperative diagnostic accuracy is at the center of making these less burdensome and more cost-effective advanced, highly targeted endoscopic outpatient surgical procedures work.[46,81,82] Currently, the accuracy of the lumbar MRI scan report in predicting acceptable levels of clinical success with spinal decompression surgery can be raised only with the addition of other ancillary tests, such as a lidocaine-containing transforaminal epidural steroid injection.[32,83–86]

In an attempt to improve the diagnostic value of the lumbar MRI scan, a uniform nomenclature of a herniated disc and spinal stenosis was proposed and published.[87] Clear definitions of bulging or herniated disc or disc protrusions were given to avoid the interchangeable and indiscriminate use of these terms without attention to detail or their clinical relevance. Several radiographic classification systems of lumbar spinal stenosis in the central canal, lateral recess, or neuroforamina have been published that clearly delineate the image-based criteria for neural element compression.[33–35,87,88] However, the dichotomy between radiological assessment of painful spine conditions and successful clinical protocols continues because current MRI reporting mainly reduces spine pain to only the assessment of mechanical encroachment of neural elements, instability, and degeneration of the intervertebral disc or facet joints.[36] Any other of the many documented and validated additional lumbar pain generators that arise from inflammation, scarring, adhesions, or tethering of spinal nerves are typically not accounted for.[38,89,90] This lack of detail in the conventional MRI reporting provided by the radiologist and how it relates to the relevant clinical context motivated us to go beyond traditional subjective visual image interpretation and prompted us to look into the deployment of modern AI to reduce the waste and improve patient outcomes in modern spine care. Therefore, we investigated the feasibility of using deep neural network self-learning

algorithms to provide written reports that not only could be increasingly more accurate and consistent with traditional verbal reads produced by a radiologist but would improve upon the current industry standards.

The results of this feasibility study of 3560 lumbar MRI scans and 17,800 levels shows that the model report generated by the network models (the RadBot) was capable of producing a verbally spelled-out report by a specific lumbar spinal level comparable in detail as to what is typically seen by the radiologist in terms of detail and scope. From training using 14,240 disc locations across 25 epochs and validating against 3560 disc locations, we observed that the accuracy, sensitivity, and specificity metrics are consistently higher than 85% for the central canal compression detectors. Despite training against double the number of neural foramina (left and right nerve roots), the ability of the model to accurately match foraminal stenosis detection to those from radiologist reads is less than that of central canal compression. This can be due to a few possible reasons—due to the more complex 3D volumetric shape of the neuroforamina and of the nerve roots as compared with the posterior section of the disc and central canal, which are much larger in comparison. Variability in the training reads resulting in differences with the stenosis indicators radiologists use to describe the diagnosis of neural foraminal stenosis may also be another limitation of the RadBot in its current modeling algorithm that we may wish to overcome with additional training of the RadBot if a clinical need arises. The overall accuracy of 81% with the MRI reporting may have several explanations. Still, the most obvious one is that the deep learning network model's accuracy may not exceed 81% with the current manually segmented 3D models. Redefining this manual segmentation to common clinically relevant painful entities of the lumbar spine may improve the accuracy. Another problem that may reside in the underlying DICOM datasets, often obtained on 1.5-T scanners, which are too noisy. Moreover, an 81% accuracy for across-the-board RadBot reading of lumbar spine MRI scans obtained in our feasibility study is approximately 15% higher than the published interobserver and intraobserver reliability rates obtained on routine reports provided by radiologists on the same scan.[91–99] Despite these limitations, the RadBot was able to give the printed MRI report in approximately 8 to 10 minutes, which in today's health care cost-savings context may save time and prevent overuse owing to improved reporting standards. Future studies will have to demonstrate the reliability of the RadBot readings with κ analysis of agreement between the RadBot and the MRI reports provided by a radiologist.

## CONCLUSIONS

We demonstrated the feasibility of using deep learning algorithms for routine reporting in spine MRI. We found the minimal disparity among accuracy, sensitivity, and specificity, which indicated first that the data were not being overfitted to the training set, and second that the frequency of false negatives and false positives were both consistent and low compared with the true positives and true negatives. In addition, variability in the training data tended to reduce overfitting and overtraining as the deep neural network models learned to focus on the common indicators and ignore differences. In future studies, we will focus on providing RadBot reliability data in correlation with painful entities in patients with spinal injuries and degenerative conditions of the lumbar spine, with the ultimate objective of improving its accuracy and predictive value of favorable clinical outcomes with intervention.

## REFERENCES

1. Larson DB, Boland GW. Imaging quality control in the era of artificial intelligence. *J Am Coll Radiol*. 2019;16:1259–1266.

2. Dito WR. Impact of new technology on laboratory informatics. bar codes, voice recognition, artificial intelligence, and CD ROMS. *Clin Lab Manage Rev*. 1990;4:114–117.

3. Gowda R, Poojary BV, Sharma M, et al. Artificial intelligence based facial recognition for mood charting among men on life style modification and it's correlation with cortisol. *Asian J Psychiatr*. 2019;43:101–104.

4. Rohr K. Fundamental limits in 3D landmark localization. *Inf Process Med Imaging*. 2005;19:286–298.

5. Awad E, Levine S, Kleiman-Weiner M, et al. Drivers are blamed more than their automated cars when both make mistakes. *Nat Hum Behav*. 2020;4:134–143.

6. Ryan C, Murphy F, Mullins M. Semiautonomous vehicle risk analysis: a telematics-based anomaly detection approach. *Risk Anal*. 2019;39:1125–1140.

7. Kallioinen N, Pershina M, Zeiser J, et al. Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. *Front Psychol*. 2019;10:2415. doi:10.3389/fpsyg.2019.02415.

8. Borenstein J, Herkert JR, Miller KW. Self-driving cars

and engineering ethics: the need for a system level analysis. *Sci Eng Ethics*. 2019;25:383–398.

9. Maxmen A. Self-driving car dilemmas reveal that moral choices are not universal. *Nature*. 2018;562:469–470. doi:10.1038/d41586-018-07135-0.

10. Wang HT, Smallwood J, Mourao-Miranda J, et al. Finding the needle in a high-dimensional haystack: canonical correlation analysis for neuroscientists. *Neuroimage*. 2020:116745. doi:10.1016/j.neuroimage.2020.116745.

11. Safdar MF, Alkobaisi SS, Zahra FT. A A comparative analysis of data augmentation approaches for magnetic resonance imaging (MRI) scan images of brain tumor. *Acta Inform Med*. 2020;28:29–36.

12. Ozawa T, Ishihara S, Fujishiro M, et al. Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *Therap Adv Gastroenterol*. 2020;13:1756284820910659. doi:10.1177/1756284820910659.

13. Choi J, Shin K, Jung J, et al. Convolutional neural network technology in endoscopic imaging: artificial intelligence for endoscopy. *Clin Endosc*. 2020;53:117–126.

14. Lee JH, Han IH, Kim DH, et al. Spine computed tomography to magnetic resonance image synthesis using generative adversarial networks : a preliminary study. *J Korean Neurosurg Soc*. 2020. doi:10.3340/jkns.2019.0084.

15. Pan Y, Chen Q, Chen T, et al. Evaluation of a computer-aided method for measuring the Cobb angle on chest X-rays. *Eur Spine J*. 2019;28:3035–3043.

16. Hopkins BS, Weber KA, II, Kesavabhotla K, et al. Machine learning for the prediction of cervical spondylotic myelopathy: a post hoc pilot study of 28 participants. *World Neurosurg*. 2019;127:e436–e442. doi:10.1016/j.wneu.2019.03.165.

17. Zhang Q, Bhalerao A, Hutchinson C. Deformable appearance pyramids for anatomy representation, landmark detection and pathology classification. *Int J Comput Assist Radiol Surg*. 2017;12:1271–1280.

18. Oktay AB, Albayrak NB, Akgul YS. Computer aided diagnosis of degenerative intervertebral disc diseases from lumbar MR images. *Comput Med Imaging Graph*. 2014;38:613–619.

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

20. Oktay AB, Akgul YS. Localization of the lumbar discs using machine learning and exact probabilistic inference. *Med Image Comput Comput Assist Interv*. 2011;14:158–165.

21. Koh J, Kim T, Chaudhary V, et al. Automatic segmentation of the spinal cord and the dural sac in lumbar MR images using gradient vector flow field. *Conf Proc IEEE Eng Med Biol Soc*. 2010;2010:3117–3120.

22. Guyer R, Musacchio M, Cammisa FP, Jr., et al. ISASS recommendations/coverage criteria for decompression with interlaminar stabilization - coverage indications, limitations, and/or medical necessity. *Int J Spine Surg*. 2016;10:41. doi:10.14444/3041

23. Chen PG, Daubs MD, Berven S, et al. Surgery for degenerative lumbar scoliosis: the development of appropriateness criteria. *Spine (Phila Pa 1976)*. 2016;41:910–918.

24. Manchikanti L, Helm Ii S, Singh V, et al. Accountable interventional pain management: a collaboration among practitioners, patients, payers, and government. *Pain Physician*. 2013;16:E635–E670.

25. Cheng JS, Lee MJ, Massicotte E, et al. Clinical guidelines and payer policies on fusion for the treatment of chronic low back pain. *Spine (Phila Pa 1976)*. 2011;36:S144–S163.

26. Trigg SD, Devilbiss Z. Spine conditions: lumbar spinal stenosis. *FP Essent*. 2017;461:21–25.

27. Expert Panel on Neurological I, McDonald MA, Kirsch CFE, et al. ACR Appropriateness Criteria® cervical neck pain or cervical radiculopathy. *J Am Coll Radiol*. 2019;16:S57–S76.

28. Hu X, Chen M, Pan J, et al. Is it appropriate to measure age-related lumbar disc degeneration on the mid-sagittal MR image? A quantitative image study. *Eur Spine J*. 2018;27:1073–1081.

29. Zarrabian M, Bidos A, Fanti C, et al. Improving spine surgical access, appropriateness and efficiency in metropolitan, urban and rural settings. *Can J Surg*. 2017;60:342–348.

30. Roth CJ, Angevine PD, Aulino JM, et al. ACR appropriateness criteria myelopathy. *J Am Coll Radiol*. 2016;13:38–44.

31. Anselmetti GC, Bernard J, Blattert T, et al. Criteria for the appropriate treatment of osteoporotic vertebral compression fractures. *Pain Physician*. 2013;16:E519–E530.

32. 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.

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

34. 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;16:431–437.

35. Pfirrmann CW, Metzdorf A, Zanetti M, et al. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine (Phila Pa 1976)*. 2001;26:1873–1878.

36. 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.

37. Yeung AT, Gore S. In-vivo endoscopic visualization of patho-anatomy in symptomatic degenerative conditions of the lumbar spine ii: intradiscal, foraminal, and central canal decompression. *Surg Technol Int*. 2011;21:299–319.

38. Yeung A, Lewandrowski K-U. Early and staged endoscopic management of common pain generators in the spine. *J Spine Surg*. 2019:S1–S5.

39. Yeung A, Roberts A, Zhu L, et al. Treatment of soft tissue and bony spinal stenosis by a visualized endoscopic transforaminal technique under local anesthesia. *Neurospine*. 2019;16:52–62.

40. Lewandrowski KU, de Carvalho PST, Calderaro AL, et al. Outcomes with transforaminal endoscopic versus percutaneous laser decompression for contained lumbar herniated

disc: a survival analysis of treatment benefit. *J Spine Surg*. 2020;6:S84–S99.

41. Lewandrowski KU, Zhang X, Ramirez Leon JF, et al. Lumbar vacuum disc, vertical instability, standalone endoscopic interbody fusion, and other treatments: an opinion based survey among minimally invasive spinal surgeons. *J Spine Surg*. 2020;6:S165–S178.

42. Yeung A, Lewandrowski KU. Five-year clinical outcomes with endoscopic transforaminal foraminoplasty for symptomatic degenerative conditions of the lumbar spine: a comparative study of inside-out versus outside-in techniques. *J Spine Surg*. 2020;6:S66–S83.

43. Yeung A, Lewandrowski KU. Early and staged endoscopic management of common pain generators in the spine. *J Spine Surg*. 2020;6:S1–S5.

44. Nicholson T, Maltenfort M, Getz C, et al. Multimodal pain management protocol versus patient controlled narcotic analgesia for postoperative pain control after shoulder arthroplasty. *Arch Bone Joint Surg*. 2018;6:196–202.

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

46. 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:S100–S115.

47. Yeung A, Wei SH. Surgical outcome of workman's comp patients undergoing endoscopic foraminal decompression for lumbar herniated disc. *J Spine Surg*. 2020;6:S116–S119.

48. Faciszewski T. Spine policy. What's in a name? *Spine J*. 2001;1:300.

49. Thorsteinsdottir B, Beck A, Tilburt JC. Grow a spine, have a heart: responding to patient requests for marginally beneficial care. *AMA J Ethics*. 2015;17:1028–1034.

50. Inglis T, Schouten R, Dalzell K, et al. Access to orthopaedic spinal specialists in the Canterbury public health system: quantifying the unmet need. *N Z Med J*. 2016;129:19–24.

51. Mok JM, Martinez M, Smith HE, et al. Impact of a bundled payment system on resource utilization during spine surgery. *Int J Spine Surg*. 2016;10:19. doi:10.14444/3019.

52. Weir S, Samnaliev M, Kuo TC, et al. Persistent postoperative pain and healthcare costs associated with instrumented and non-instrumented spinal surgery: a case-control study. *J Orthop Surg Res*. 2020;15:127. doi:10.1186/s13018-020-01633-6.

53. Marrache M, Harris AB, Raad M, et al. Pre-operative and post-operative spending among working age adults undergoing posterior spinal fusion surgery for degenerative pathology. *World Neurosurg*. 2020. doi:10.1016/j.wneu.2020.03.143.

54. Jain A, Yeramaneni S, Kebaish KM, et al. Cost-utility analysis of rhBMP-2 use in adult spinal deformity surgery. *Spine (Phila Pa 1976)*. 2020. doi:10.1097/BRS.0000000000003442.

55. Safaee MM, Dalle Ore CL, Zygourakis CC, et al. Estimating a price point for cost-benefit of bone morphogenetic protein in pseudarthrosis prevention for adult spinal deformity surgery. *J Neurosurg Spine*. 2019:1–8.

56. Jonsson E, Hansson-Hedblom A, Kirketeig T, et al. Cost and health outcomes patterns in patients treated with spinal cord stimulation following spine surgery-a register-based study. *Neuromodulation*. 2019. doi:10.1111/ner.13056.

57. Hansson-Hedblom A, Jonsson E, Fritzell P, et al. The association between patient reported outcomes of spinal surgery and societal costs: a register based study. *Spine (Phila Pa 1976)*. 2019;44:1309–1317.

58. Carr DA, Saigal R, Zhang F, et al. Enhanced perioperative care and decreased cost and length of stay after elective major spinal surgery. *Neurosurg Focus*. 2019;46:E5. doi:10.3171/2019.1.FOCUS18630.

59. Ball JR, Sekhon LH. Timing of decompression and fixation after spinal cord injury–when is surgery optimal? *Crit Care Resusc*. 2006;8:56–63.

60. Whitmore RG, Stephen J, Stein SC, et al. Patient comorbidities and complications after spinal surgery: a societal-based cost analysis. *Spine (Phila Pa 1976)*. 2012;37:1065–1071.

61. Ciol MA, Deyo RA, Howell E, et al. An assessment of surgery for spinal stenosis: time trends, geographic variations, complications, and reoperations. *J Am Geriatr Soc*. 1996;44:285–290.

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

63. 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;158:807–811.

64. Yuan S, Zou Y, Li Y, et al. A clinically relevant MRI grading system for lumbar central canal stenosis. *Clin Imaging*. 2016;40:1140–1145.

65. Lee GY, Lee JW, Choi HS, et al. A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method. *Skeletal Radiol*. 2011;40:1033–1039.

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

67. Lu T, Pedemonte S, Bizzo B, et al. DeepSPINE: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. *Comput Sci*. 2018. Preprint posted online July 26, 2018. arXiv:1807.10215 [cs.CV]

68. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J Machine Learn Res*. 2006:2399–2434.

69. Chapelle O, Schölkopf B, Zien A. Semi-Supervised Learning. Q325.75(S42)(2006).

70. Oliver A, Odena A, Raffel C, et al. Realistic evaluation of deep semi-supervised learning algorithms. *Comput Sci*. Preprint posted online April 24, 2018. arXiv:1804.09170 [cs.LG]

71. Hyun M, Jeong J and Kwak N. Class-imbalanced semi-supervised learning. *Comput Sci*. Preprint posted online February 17, 2020. arXiv:2002.06815 [cs.LG]

72. Jégou S, Drozdzal M, Vazquez D, et al. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. *Comput Sci*. Preprint posted online November 28, 2016. arXiv:1611.09326 [cs.CV]

73. Simonyan K, Zisserman A. Very deep convolutional

networks for large-scale image recognition. *Comput Sci*. Preprint posted online April 10, 2015. arXiv:1409.1556v6 [cs.CV]

74. Kreiner DS, Baisden J, Gilbert T, et al. Re: diagnostic tests the NASS stenosis guidelines. *Spine J*. 2014;14:201–202.

75. Haig AJ. Diagnostic tests the NASS stenosis guidelines. *Spine J*. 2014;14:200–201.

76. Jacobs E, Senden R, McCrum C, et al. Effect of a semirigid thoracolumbar orthosis on gait and sagittal alignment in patients with an osteoporotic vertebral compression fracture. *Clin Interv Aging*. 2019;14:671–680.

77. Boden SD, Davis DO, Dina TS, et al. Abnormal magnetic-resonance scans of the lumbar spine in asymptomatic subjects. A prospective investigation. *J Bone Joint Surg Am*. 1990;72:403–408.

78. 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.

79. Lewandrowski KU. The strategies behind "inside-out" and "outside-in" endoscopy of the lumbar spine: treating the pain generator. *J Spine Surg*. 2020;6:S35–S39. doi:10.21037/jss.2019.06.06.

80. Lewandrowski KU, Yeung A. Meaningful outcome research to validate endoscopic treatment of common lumbar pain generators with durability analysis. *J Spine Surg*. 2020;6:S6–S13. doi:10.21037/jss.2019.09.07.

81. Fujii Y, Yamashita K, Sugiura K, et al. Early return to activity after minimally invasive full endoscopic decompression surgery in medical doctors. *Journal of Spine Surgery*. 2019:S294–S299.

82. Maeda T, Takamatsu N, Hashimoto A, et al. Return to play in professional baseball players following transforaminal endoscopic decompressive spine surgery under local anesthesia. *Journal of Spine Surgery*. 2020:S300–S306.

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

84. Lee JW, Kim SH, Lee IS, et al. Therapeutic effect and outcome predictors of sciatica treated using transforaminal epidural steroid injection. *AJR Am J Roentgenol*. 2006;187:1427–1431.

85. 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;28:204–208.

86. Ercalik T, Gencer Atalay K, Sanal Toprak C, et al. Outcome measurement in patients with low back pain undergoing epidural steroid injection. *Turk J Phys Med Rehabil*. 2019;65:154–159.

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

88. Hasegawa T, An HS, Haughton VM, et al. Lumbar foraminal stenosis: critical heights of the intervertebral discs and foramina. A cryomicrotome study in cadavera. *J Bone Joint Surg Am*. 1995;77:32–38.

89. Gore S, Yeung A. The "inside out" transforaminal technique to treat lumbar spinal pain in an awake and aware patient under local anesthesia: results and a review of the literature. *Int J Spine Surg*. 2014;8. doi:10.14444/1028.

90. Tsou PM, Alan Yeung C, Yeung AT. Posterolateral transforaminal selective endoscopic discectomy and thermal annuloplasty for chronic lumbar discogenic pain: a minimal access visualized intradiscal surgical procedure. *Spine J*. 2004;4:564–573.

91. Gupta A, Upadhyaya S, Yeung CM, et al. Disk area is a more reliable measurement than anteroposterior length in the assessment of lumbar disk herniations: a validation study. *Clin Spine Surg*. 2020. doi:10.1097/BSD.0000000000000958.

92. Schroeder GD, Suleiman LI, Chioffe MA, et al. The effect of oblique magnetic resonance imaging on surgical decision making for patients undergoing an anterior cervical discectomy and fusion for cervical radiculopathy. *Int J Spine Surg*. 2019;13:302–307. doi:10.14444/6041.

93. Li Y, Fredrickson V, Resnick DK. How should we grade lumbar disc herniation and nerve root compression? A systematic review. *Clin Orthop Relat Res*. 2015;473:1896–1902.

94. Fyllos AH, Arvanitis DL, Karantanas AH, et al. Magnetic resonance morphometry of the adult normal lumbar intervertebral space. *Surg Radiol Anat*. 2018;40:1055–1061.

95. Lonne G, Odegard B, Johnsen LG, et al. MRI evaluation of lumbar spinal stenosis: is a rapid visual assessment as good as area measurement? *Eur Spine J*. 2014;23:1320–1324. 2014/02/28. doi:10.1007/s00586-014-3248-4.

96. Sher I, Daly C, Oehme D, et al. Novel application of the pfirrmann disc degeneration grading system to 9.4T MRI: higher reliability compared to 3T MRI. *Spine (Phila Pa 1976)*. 2019;44:E766–E773. doi:10.1097/BRS.0000000000002967.

97. Little JW, Grieve T, Cantu J, et al. Reliability of human lumbar facet joint degeneration severity assessed by magnetic resonance imaging. *J Manipulative Physiol Ther*. 2020. doi:10.1016/j.jmpt.2018.11.027.

98. Mo AZ, Miller PE, Glotzbecker MP, et al. The reliability of the OASpine thoracolumbar classification system in children: results of a multicenter study. *J Pediatr Orthop*. 2020;40:e352–e356. doi:10.1097/BPO.0000000000001521.

99. Battaglia PJ, Maeda Y, Welk A, et al. Reliability of the Goutallier classification in quantifying muscle fatty degeneration in the lumbar multifidus using magnetic resonance imaging. *J Manipulative Physiol Ther*. 2014;37:190–197.

push any 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.