

**UCSF**

**UC San Francisco Previously Published Works**

**Title**

Evaluation of 2 Novel Ratio-Based Metrics for Lumbar Spinal Stenosis.

**Permalink**

<https://escholarship.org/uc/item/1nm793hj>

**Journal**

American Journal of Neuroradiology, 43(10)

**Authors**

Link, T  
Chin, C  
Ben-Natan, A  
[et al.](#)

**Publication Date**







2022-10-01

**DOI**

10.3174/ajnr.A7638

Peer reviewed

# Evaluation of 2 Novel Ratio-Based Metrics for Lumbar Spinal Stenosis

 U.U. Bharadwaj,  A.R. Ben-Natan,  J. Huang, V. Pedoia,  D. Chou,  S. Majumdar, T.M. Link, and  C.T. Chin



## ABSTRACT

**BACKGROUND AND PURPOSE:** Quantitative metrics of the dural sac such as the cross-sectional area are commonly used to evaluate central canal stenosis. The aim of this study was to analyze 2 new metrics to measure spinal stenosis on the basis of the ratio between the dural sac and disc cross-sectional areas (DDRCA) and the dural sac and disc anterior-posterior diameters (DDRDA) and compare them with established quantitative metrics of the dural sac.

**MATERIALS AND METHODS:** T2-weighted axial MR images ( $n = 260$  patients) were retrospectively evaluated, graded for central canal stenosis as normal (no stenosis), mild, moderate, or severe from L1/L2 through L5/S1 with 1 grade per spinal level and annotated to measure the DDRCA and DDRDA. Thresholds were obtained using a decision tree classifier on a subset of patients ( $n = 130$ ) and evaluated on the remaining patients ( $n = 130$ ) for accuracy and consistency across demographics, anatomic variation, and clinical outcomes.

**RESULTS:** DDRCA and DDRDA had areas under the receiver operating characteristic curve of 98.6 (97.4–99.3) and 98.0 (96.7–98.9) compared with dural sac cross-sectional area at 96.5 (95.0–97.7) for binary classification. DDRDA and DDRCA had  $\kappa$  scores of 0.75 (0.71–0.79) and 0.80 (0.75–0.83) compared with dural sac cross-sectional area at 0.62 (0.57–0.66) for multigrade classification. No significant differences ( $P > .1$ ) in the area under the receiver operating characteristic curve were observed for the DDRDA across variations in the body mass index. The DDRDA also had the highest area under the receiver operating characteristic curve among symptomatic patients (visual analog scale  $\geq 7$ ) or patients who underwent surgery.

**CONCLUSIONS:** Ratio-based metrics (DDRDA and DDRCA) are accurate and robust to anatomic and demographic variability compared with quantitative metrics of the dural sac and better correlated with symptomatology and surgical outcomes.

**ABBREVIATIONS:** AUROC = area under the receiver operating characteristic curve; BMI = body mass index; DDRCA = ratio between dural sac and disc cross-sectional areas; DDRDA = ratio between dural sac and disc anterior-posterior diameters; DSCA = dural sac cross-sectional area; DSDIA = dural sac anterior-posterior diameter; LSS = lumbar spinal stenosis; VAS = visual analog scale

Lumbar spinal stenosis (LSS) is one of the most common causes for lumbar spinal surgery in patients older than 65 years of age.<sup>1</sup> The etiology is multifactorial but predominantly attributed to degenerative changes. Degenerative canal narrowing can be secondary

to changes that include disc protrusion, extrusion; ligamentum flavum hypertrophy; or facet joint arthropathy.<sup>2</sup> Historically, radiographic LSS has been described using morphologic categories ranging from any narrowing of the spinal canal<sup>3</sup> to more detailed descriptors evaluating CSF space obliteration and neural element separation;<sup>4</sup> nevertheless, classification of LSS is highly variable, with a number of grading systems, none of which are widely accepted.<sup>5</sup>

Accurate classification of LSS, however, is essential for subsequent patient management.<sup>6</sup> Clinical symptoms and examination and radiologic findings are all integral and contribute to the diagnosis of symptomatic LSS. There are no physical examination findings or clinical history that is both highly sensitive and specific for diagnosing LSS;<sup>7</sup> imaging can, therefore, confirm the structural diagnosis and clarify the anatomy if therapeutic management such as injections or surgery is contemplated. When imaging is indicated,


Received January 30, 2022; accepted after revision July 25.


From the Departments of Radiology and Biomedical Imaging (U.U.B., V.P., S.M., T.M.L., C.T.C.) and Neurological Surgery (A.R.B.-N., J.H., D.C.), University of California San Francisco, San Francisco, California.

A.R. Ben-Natan and J. Huang contributed equally to this work.

This work was funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases, grant/award No: UH2AR076724-01.

Please address correspondence to Upasana Upadhyay Bharadwaj, MD, Department of Radiology and Biomedical Imaging, University of California San Francisco, 185 Berry St, Suite 350, San Francisco, CA 94107; e-mail: Upasana.Bharadwaj@ucsf.edu; @UUBharad

 Indicates open access to non-subscribers at [www.ajnr.org](http://www.ajnr.org)

 Indicates article with online supplemental data.

<http://dx.doi.org/10.3174/ajnr.A7638>

MR imaging is widely accepted as the preferred technique owing to its superior soft-tissue contrast<sup>8,9</sup> and various qualitative, morphologic features; quantitative metrics have been proposed for LSS on MR imaging.<sup>4,10,11</sup>

To optimize the effects of variability, poor agreement, and sub-optimal outcomes associated with qualitative features,<sup>12</sup> articles in the literature have proposed quantitative measures for diagnosing and grading LSS.<sup>13,14</sup> The anterior-posterior diameter of the dural sac (DSDIA) and the dural sac cross-sectional area (DSCA) have been evaluated extensively in prior studies with limited success in establishing clinical utility;<sup>15-19</sup> moreover, various thresholds have been proposed for each measure.<sup>14,20</sup> A DSCA of  $<100 \text{ mm}^2$  at more than 2 of 3 intervertebral levels (L2/L3, L3/L4, L4/L5) was shown to be highly associated with the presence of intermittent claudication;<sup>17</sup> and pronounced stenosis of the canal (DSDIA of  $<6 \text{ mm}$  on myelography) predicted less postoperative pain in a 5-year follow-up study.<sup>21</sup> The increasing number of quantitative measures and potential correlations with outcomes can lead to confusion in the clinical routine because even specialized radiologists apply each measure differently<sup>22-24</sup> according to the results of a recent Delphi survey.<sup>25</sup>

Furthermore, a weakness of commonly used nonratio metrics such as DSCA and DSDIA is that they are not anatomically normalized and incorporate only the absolute distance or area, possibly explaining the high variability and susceptibility to demographic changes.

Given the wide variability of the quantitative measurements and correlation with symptoms and outcomes, a reproducible quantitative grading system for LSS is essential for subsequent management. In this study, we propose to calculate ratios measured at the disc level, the most stenotic level, relative to the dural sac: the dural sac-to-disc ratio of the respective anterior-posterior diameters (DDRDI) and the dural sac-to-disc ratio of the respective cross-sectional areas (DDRCA) as normalized quantitative metrics for classifying stenosis. We hypothesize that these ratios incorporating the disc level may be better correlated with symptomatology and surgical outcomes compared with quantitative metrics of the dural sac.

## MATERIALS AND METHODS

### Study Design

In this institutional review board–approved retrospective cross-sectional study, lumbar spine MRIs along with clinical data were evaluated to assess our proposed quantitative metrics, DDRDI and DDRCA, for grading LSS and comparing it with other more commonly used nonratio metrics such as the DSCA (standard of reference) as well as DSDIA.

### Patient Cohort

Patients who underwent lumbar spine MR imaging for clinical indications between 2008 and 2019 were included after applying the following exclusion criteria: Those with age younger than 19 years, transitional anatomy, fractures, postoperative changes, extensive hardware, infection, primary tumors, and widespread metastatic disease to the spine were excluded. Studies with the absence of a T2-weighted axial sequence or poor image quality were also excluded. A total of 30,619 patients were identified, of whom

a subset of patients ( $n = 260$ ) were selected at random, with uniform sampling to be included in the study.

### Clinical Data

We collected the following clinical data: presenting symptoms, low back pain, and the radicular pain score on a visual analog scale (VAS),<sup>26</sup> ranging from 0 to 10; demographics including age, sex, and body mass index (BMI) from the electronic health record; as well as clinical management spanning noninvasive treatment to surgical procedures.

### Image Acquisition

All T2-weighted axial MRIs used in this study were FSE sequences acquired in our institution as part of routine clinical lumbar spine MR imaging studies using a 3T MR imaging scanner (Discovery MR750; GE Healthcare) with a section thickness of 4.0 mm, section spacing of 1.0 mm, FOV of 18.0 cm, TE of 85.0 ms, TR of 4202.0 ms, flip angle of 115°, and a matrix of  $512 \times 512$  pixels. Axial sequences were acquired in the contiguous axial plane as per the imaging protocols at our institution, with no disc-specific adjustments such as disc space–targeted angled axial images.

### Grading LSS

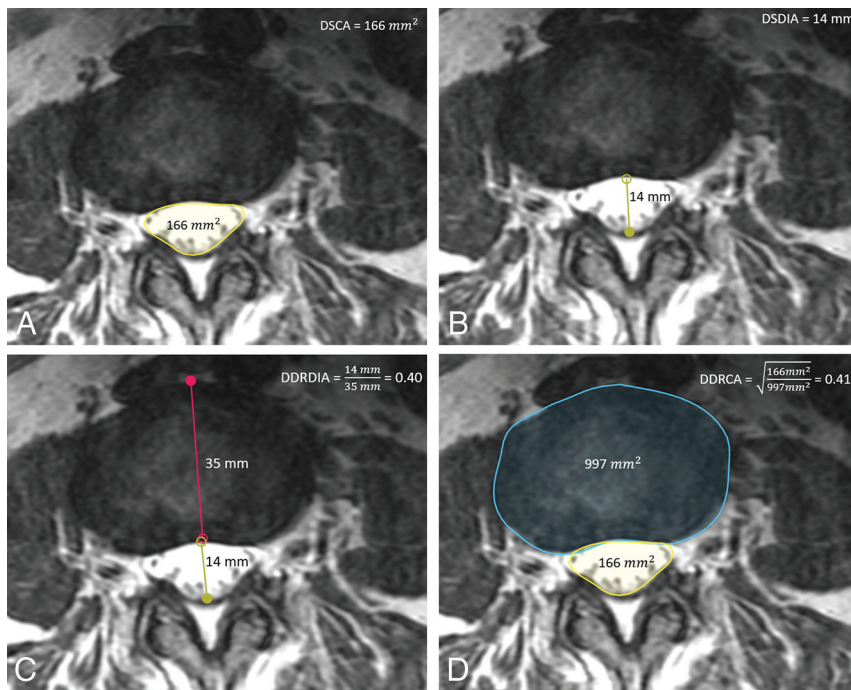
A board-certified neuroradiologist (R1) with 25 years of experience qualitatively graded MRIs from the study cohort ( $n = 260$ ) for central canal stenosis as normal (absence of stenosis), mild, moderate, or severe with 1 grade per spinal level (L1/L2, L2/L3, L3/L4, L4/L5, L5/S1). Grading was based on a published qualitative grading system (Schizas system) as follows:<sup>11</sup> Normal indicated absence of LSS based on qualitative criteria on MR imaging: homogeneous distribution of the CSF and individual rootlets visualized in the dorsal aspect of the dural sac; mild, some CSF present and the rootlets still individualized; moderate, the rootlets occupying the entire dural sac with minimal-to-no CSF, epidural fat visualized dorsally; and severe, thecal sac obliterated and no epidural fat, CSF, or individual rootlets visualized.

### Quantitative Metrics

This study evaluates the proposed metrics: DDRDI, which measures the ratio between the anterior-posterior diameters of the dural sac and intervertebral disc, and DDRCA, which measures the ratio between the cross-section areas of the dural sac and intervertebral disc as well as standard metrics such as DSCA and DSDIA. Using a research annotation platform (MD.ai; md.ai/), a trained researcher and a radiology trainee (R2) annotated the T2-weighted axial slices with free-form masks of the dural sac and intervertebral disc, as well as lines for measuring their respective anterior-posterior diameters from which the metrics were computed, as shown in Fig 1. Figure 2 provides examples of normal (no stenosis), mild, moderate, and severe stenosis with metric values.

### Cutoff Thresholds for Grading Stenosis

The study cohort ( $n = 260$ ) was partitioned randomly into 2 distinct groups: 1) a development cohort ( $n = 130$ ), used to determine thresholds for DDRDI, DDRCA, DSDIA, and DSCA (standard of reference); and 2) an evaluation cohort ( $n = 130$ ) in which all metrics were evaluated.



**FIG 1.** Sample T2-weighted axial section at L2/L3 graded normal with the following: A, Free-form annotation around the DSCA of 166 mm<sup>2</sup>. B, Line annotation with a DSDIA of 14 mm. C, Line annotations with a DDRDIA of 0.4. D, Free-form annotations with DDRCA of 0.41. The square root is used as a normalization step to account for the quadraticity of area measures.

For each metric, a decision tree classifier was fit on the development cohort ( $n = 130$ ) using R1's grades as ground truth to determine cutoff thresholds for classifying a given T2-weighted axial section as having normal (no stenosis), mild, moderate, or severe stenosis. The decision tree is a statistical modelling technique that automatically creates branches of decisions based on each measurement and its corresponding ground truth grade so that the total classification error is minimized.<sup>27</sup> Decision trees have been previously used to obtain thresholds for LSS and offer the advantage of clinically interpretable rules.<sup>28</sup> The Scikit-learn Python library, Version 0.24.2 DecisionTreeClassifier module (<https://scikit-learn.org/stable/index.html>) was used with the `max_depth` parameter set to 3 and `max_leaves` set to 4 to avoid overfitting.<sup>29</sup>

### Statistical Analysis

All analyses were performed on the evaluation cohort ( $n = 130$ ). Statistical power analysis for pair-wise comparison of the quantitative metrics with an assumed effect size of 0.55,  $\alpha$  of .05,  $\beta$  of 0.2, and power of 80% resulted in a minimum sample size of 120. The SciPy Version 1.6.0 Python library and its stats module were used for all statistical analyses reported in this article.<sup>30</sup>

**Association with Stenosis.** The decision tree classifiers fit on the development cohort ( $n = 130$ ) were used to classify 1 section from each disc level of the evaluation cohort ( $n = 130$ ) as normal, mild, moderate, or severe. Association with stenosis for each metric was characterized for both binarized grading of stenosis (normal/mild versus moderate/severe) and multigrade classification.

Binary classification was evaluated using the area under the receiver operation characteristic curve (AUROC). Statistical significance of pair-wise differences in the AUROC corresponding to each quantitative metric was characterized using the DeLong test for comparing AUROCs, with  $P < .05$  considered statistically significant.<sup>31</sup> Evaluation was bootstrapped to generate 95% confidence intervals.

Association of each metric with stenosis in the multigrade setting was evaluated using model accuracy, multi-class AUROC with the one-vs-one criterion, and agreement with R1's grades using a linearly-weighted Cohen  $\kappa$  coefficient.

**Demographic Variability.** The AUROC for binarized grading of stenosis using each metric as a score was used to assess consistency across demographics. AUROC values were computed for sex splits (male versus female), age splits using 45 years as a cutoff (age younger than 45 years versus age 45 years or older),<sup>32</sup> and BMI splits using a mean

BMI of 25.0 kg/m<sup>2</sup> as a cutoff (BMI < 25.0 kg/m<sup>2</sup> versus BMI  $\geq$  25.0 kg/m<sup>2</sup>).<sup>33</sup>

**Symptomatology.** The AUROC for binarized grading of stenosis was used to assess the accuracy of each metric across 2 groups: VAS < 7 and VAS  $\geq$  7.

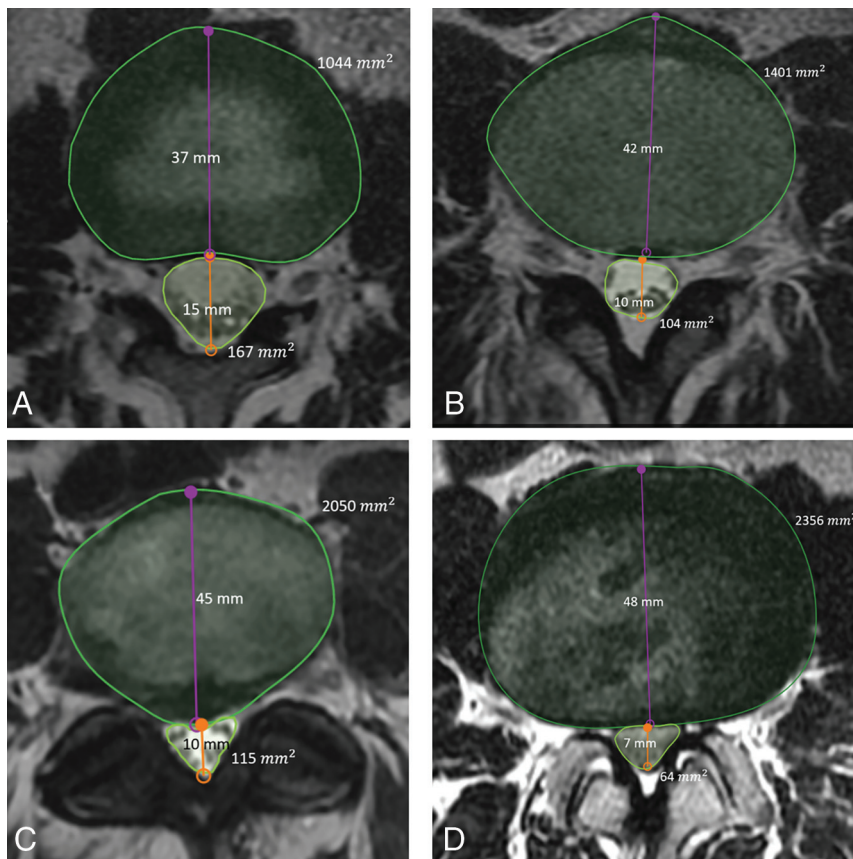
**Anatomic Normalization.** To evaluate anatomic normalization of each quantitative metric, we clustered spinal levels into 2 groups: 1) upper lumbar levels consisting of L1/L2, L2/L3, and L3/L4, and 2) lower lumbar levels consisting of L4/L5 and L5/S1.

**Association with Prognosis.** The utility of each metric in association with outcomes was assessed on a subset of the evaluation cohort ( $n = 130$ ), referred to as the "prognostic cohort" ( $n = 58$ ) with patients managed conservatively ( $n = 30$ ) and those who went on to require surgery ( $n = 28$ ). Using each metric as a score, we evaluated the AUROC associated with predicting surgery from the preoperative MR imaging. Only the symptomatic level or the level at which surgery was performed was included in this analysis.

The decision tree classifier was used to grade each spinal level of the prognostic cohort ( $n = 58$ ) as normal, mild, moderate, or severe. Linearly-weighted Cohen  $\kappa$  scores were computed for surgical-versus-conservatively managed cases.

**Agreement and Reproducibility.** To characterize reliability for the proposed metrics, R2 and R1 annotated another subset of patients ( $n = 40$ ) for lines from the evaluation cohort ( $n = 130$ ).





**FIG 2.** Sample T2-weighted axial MR imaging slices of the lumbar spine for each stenosis grade, determined qualitatively by a neuroradiologist, with the metrics annotated. A, Grade: normal; level, L1/L2; DSCA, 167 mm<sup>2</sup>; DSDIA, 15 mm; DDRDIA, 0.42; DDRCA, 0.40. B, Grade: mild; level, L3/L4; DSCA, 104 mm<sup>2</sup>; DSDIA, 10 mm; DDRDIA, 0.24; DDRCA, 0.27. C, Grade: moderate; level L4/L5; DSCA, 115 mm<sup>2</sup>; DSDIA, 10 mm; DDRDIA, 0.22; DDRCA, 0.24. D, Grade: severe; level, L2/L3; DSCA, 64 mm<sup>2</sup>; DSDIA, 7 mm; DDRDIA, 0.14; DDRCA, 0.16.

Reproducibility of estimating DDRDIA was computed using the concordance correlation coefficient.

To characterize interrater agreement for the qualitative grading of lumbar spinal stenosis, R1, R2, and a board-certified musculoskeletal radiologist (R3) with 23 years of experience assessed another subset of patients ( $n = 32$ ) from the evaluation cohort ( $n = 130$ ). Interrater agreement among R1, R2, and R3 was evaluated using a linearly-weighted Cohen  $\kappa$  coefficient.

## RESULTS

### Patient Cohort

The development cohort ( $n = 130$ ) consisted of 65 female and 65 male patients, with a mean age of 57.6 (20.0–96.0) years and a mean BMI of 26.9 (15.3–58.8) kg/m<sup>2</sup>. Patients presented with either low back pain ( $n = 33$ ), radicular pain ( $n = 14$ ), or both low back pain and radicular pain ( $n = 68$ ), as well as other symptoms ( $n = 15$ ) including numbness, tingling, weakness, dysesthesia, and tightness. Patients in the development cohort had an average low back pain score of 5.8 (SD, 2.6) and a radicular pain score of 5.9 (SD, 2.7) on an 11-point qualitative numeric pain rating scale.

The evaluation cohort ( $n = 130$ ) consisted of 58 female and 72 male patients with a mean age of 58.3 (19.0–84.0) years and a mean BMI of 26.7 (17.5–41.3) kg/m<sup>2</sup>. Patients in this cohort presented with low back pain ( $n = 27$ ), radicular pain ( $n = 20$ ), both ( $n = 72$ ), and other symptoms ( $n = 11$ ) including numbness, weakness, and tightness. Patients in the evaluation cohort had an average low back pain score of 5.8 (SD, 2.4) and a radicular pain score of 6.0 (SD, 2.5) on the numeric rating scale.

### LSS Grades

A total of 555 slices were graded in the development cohort with the following distribution: normal ( $n = 273$ , 49.2%), mild ( $n = 200$ , 36.0%), moderate ( $n = 45$ , 8.1%), and severe ( $n = 37$ , 6.7%) stenosis across lumbar spinal levels L1/L2 ( $n = 113$ , 20.4%), L2/L3 ( $n = 121$ , 21.8%), L3/L4 ( $n = 122$ , 21.9%), L4/L5 ( $n = 113$ , 20.4%), and L5/S1 ( $n = 86$ , 15.5%).

A total of 491 slices were graded in the evaluation cohort with the following distribution: normal ( $n = 244$ , 49.7%), mild ( $n = 149$ , 30.3%), moderate ( $n = 37$ , 7.5%), and severe ( $n = 61$ , 12.5%) stenosis across lumbar spinal levels L1/L2 ( $n = 111$ , 22.6%), L2/L3 ( $n = 113$ , 23.0%), L3/L4 ( $n = 108$ , 22.0%), L4/L5 ( $n = 96$ , 19.6%), and L5/S1 ( $n = 63$ , 12.8%).

### Cutoff Thresholds for Grading Stenosis

The decision tree for each quantitative metric was of depth 3 as visualized in Fig 3. Cutoff thresholds for grading stenosis using each metric were derived as follows:

DDRDIA: normal, DDRDIA  $\geq 0.36$ ; mild,  $0.24 \leq$  DDRDIA  $< 0.36$ ; moderate,  $0.15 \leq$  DDRDIA  $< 0.24$ ; severe, DDRDIA  $< 0.15$ .

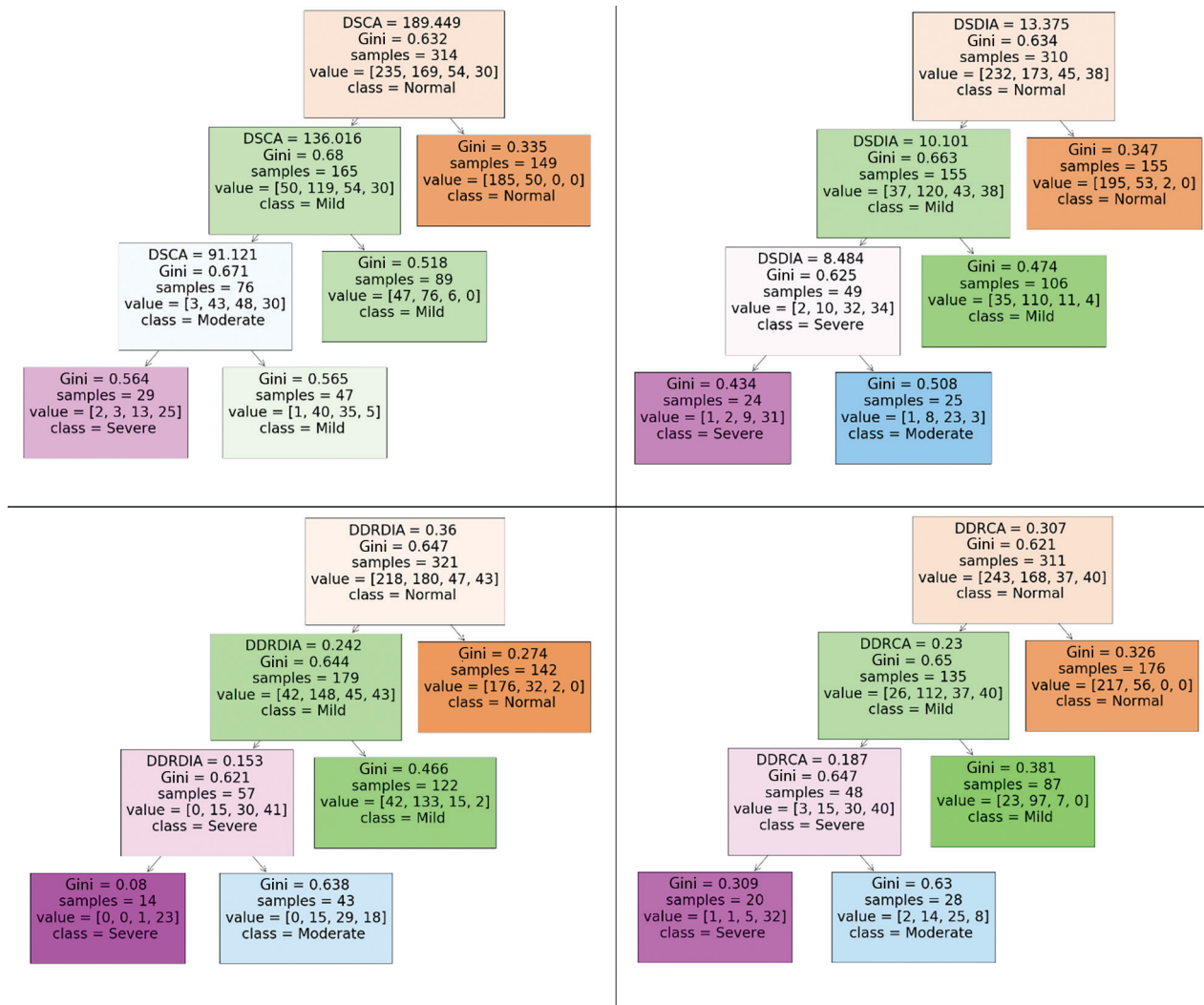
DDRCA: normal, DDRCA  $\geq 0.31$ ; mild,  $0.23 \leq$  DDRCA  $< 0.31$ ; moderate,  $0.19 \leq$  DDRCA  $< 0.23$ ; severe, DDRCA  $< 0.19$ .

DSCA: normal, DSCA  $\geq 189.5$  mm<sup>2</sup>; mild,  $136.0$  mm<sup>2</sup>  $\leq$  DSCA  $< 189.5$  mm<sup>2</sup>; moderate,  $91.1$  mm<sup>2</sup>  $\leq$  DSCA  $< 136.0$  mm<sup>2</sup>; severe, DSCA  $< 91.1$  mm<sup>2</sup>.

DSDIA: normal, DSDIA  $\geq 13.4$  mm; mild,  $10.1$  mm  $\leq$  DSDIA  $< 13.4$  mm; moderate,  $8.5$  mm  $\leq$  DSDIA  $< 10.1$  mm; severe, DSDIA  $< 8.5$  mm.

### Statistical Analysis

**Association with Stenosis.** The proposed metrics, DDRCA and DDRDIA, had the highest AUROC for binarized classification of stenosis at 98.6 (97.4–99.3) and 98.0 (96.7–98.9), respectively, which were significantly higher ( $P < .05$ ) than the standard of reference metrics, DSCA and DSDIA, with an AUROC of 96.5



**FIG 3.** Decision rules and cutoff thresholds generated by a decision tree classifier (maximum depth = 3, maximum leaves = 4, criterion = Gini impurity) for each quantitative metric.

(95.0–97.7) and DSDIA at 96.6 (95.1–97.8). The results are presented in Table 1 and visualized in Fig 4.

DDRCA and DDRDIA had the highest agreement with R1 for multigrade classification of stenosis with  $\kappa$  values of 0.80 (0.75–0.83) and 0.75 (0.71–0.79), respectively, compared with DSCA at 0.62 (0.57–0.66) and DSDIA at 0.69 (0.64–0.75), respectively. Multiclass accuracy, AUROC, and  $\kappa$  scores for each metric are presented in Table 2.

**Demographic Variability.** All 4 quantitative metrics had higher AUROC values for men compared with women ( $P < .001$ ). No significant difference ( $P < .1$ ) in the AUROC was observed in the case of the proposed metric DDRDIA across BMI groups. The other 3 metrics (DDRCA, DSDIA, DSCA) all had significant differences in the AUROC among the demographic splits ( $P < .001$ ).

**Symptomatology.** DDRDIA had a higher AUROC than all other metrics in cases with VAS  $\geq 7$ . DDRDIA was also the only metric in which the AUROC for cases with VAS  $\geq 7$  was significantly higher ( $P < .001$ ) than that of cases with VAS  $< 7$ . AUROC values are presented in Table 3.

**Anatomic Normalization.** No significant differences were observed in the values of DDRDIA and DDRCA for stenotic cases (mild, moderate, or severe) across the upper lumbar levels (L1/L2, L2/L3, L3/L4) and the lower lumbar levels (L4/L5 and L5/S1). The standard-of-reference metrics, DSCA and DSDIA, were sensitive to anatomic changes in cases with stenosis ( $P < .001$ ).

**Association with Prognosis.** The DDRCA had the highest AUROC for predicting surgery at each spinal level from the prognostic cohort ( $n = 58$ ), with a value of 83.5 (76.6–90.1), which was significantly greater than the standard-of-reference DSCA and DSDIA, which had AUROCs of 82.4 (75.5–90.4) and 81.3 (73.2–89.4). DDRDIA had the lowest AUROC for predicting surgery, with a value of 80.8 (73.0–89.5). These results are reported in Table 4.

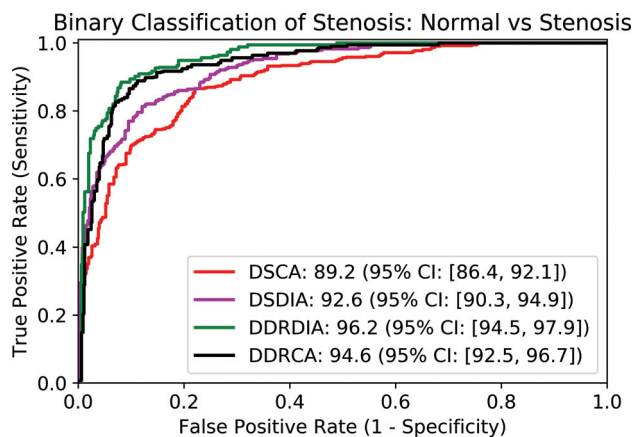
The DDRDIA had the highest agreement with R1 for multiclass grading of stenosis in surgical cases, with a  $\kappa$  coefficient of 0.77 (0.65–0.90).  $\kappa$  scores for all other metrics across surgical-versus-nonsurgical levels were significantly lower, reported in Table 5.

**Table 1: Classification of each spinal level as normal versus stenosed on the evaluation cohort (n = 130) using each quantitative metric<sup>a</sup>**

Metrics Compared	AUROC: Group 1	AUROC: Group 2	Z-Statistic	Significance
DSCA vs DSDIA	96.5 (95.0–97.7)	96.6 (95.1–97.8)	0.19	<i>P</i> = .85
DSCA vs DDRDIA	96.5 (95.0–97.7)	98.0 (96.7–98.9) <sup>b</sup>	2.24 <sup>b</sup>	<i>P</i> = .02 <sup>b</sup>
DSCA vs DDRCA	96.5 (95.0–97.7)	98.6 (97.4–99.3) <sup>b</sup>	3.54 <sup>b</sup>	<i>P</i> = .004 <sup>b</sup>
DSDIA vs DDRDIA	96.6 (95.1–97.8)	98.0 (96.7–98.9) <sup>b</sup>	2.56 <sup>b</sup>	<i>P</i> = .01 <sup>b</sup>
DSDIA vs DDRCA	96.6 (95.1–97.8)	98.6 (97.4–99.3) <sup>b</sup>	2.89 <sup>b</sup>	<i>P</i> = .004 <sup>b</sup>
DDRDIA vs DDRCA	98.0 (96.7–98.9)	98.6 (97.4–99.3)	1.57	<i>P</i> = .12

<sup>a</sup>Reported are pair-wise comparisons between the quantitative metrics using the DeLong paired test for AUROCs. Data in parentheses represent 95% confidence intervals.

<sup>b</sup>Difference in AUROC is statistically significant (*P* < .05).



**FIG 4.** Receiver operator characteristic curve using each quantitative metric as a score for binary classification of stenosis. AUROC values are reported in the legend.

**Agreement and Reproducibility.** The reproducibility of measuring DDRDIA was substantial, with a concordance correlation coefficient of 0.9 between R1 and R2. Interrater agreement among (R1, R2), (R1, R3), and (R2, R3) was substantial with  $\kappa$  scores of 0.79, 0.72, and 0.65, respectively.

## DISCUSSION

In this study, we proposed 2 ratio-based metrics for grading LSS, which, to our knowledge, has heretofore not been investigated in the literature. DDRDIA and DDRCA measured ratios between the dural sac and the intervertebral disc of the anterior-posterior diameters and cross-sectional areas, respectively. Our ratio-based approach naturally lends itself to a normalized metric between 0 and 1, which can be interpreted clinically as a surrogate for the severity of stenosis.

Our results indicate that DDRDIA and DDRCA perform as well or superior to the standard of reference metrics such as DSCA and DSDIA. Prescribed thresholds for DSCA are typically binarized into normal versus stenosed or normal/mild/moderate stenosis versus severe stenosis.<sup>13,14</sup> Our study provides more fine-grained thresholds for grading stenosis using each quantitative metric. The thresholds generated by a decision tree for DSCA and DSDIA, 91.1 mm<sup>2</sup> and 10.1 mm, respectively, are consistent with previously published values for these metrics,<sup>14</sup> further validating our methodology to obtain thresholds using a decision tree classifier.

Our analysis suggests that ratio-based metrics such as DDRDIA and DDRCA are more consistent across demographic variability,

anatomically normalized, and better correlated with symptomatology and clinical outcomes compared with nonratio metrics such as DSCA and DSDIA.

DDRCA had a linearly weighted  $\kappa$  score of 0.80 using R1's grades as the ground truth, which is higher (albeit not statistically significant) than the agreement between R1 and R2 as well as between R1 and R3. High accuracy of DDRCA is an encouraging step toward multigrade classification of stenosis using ratio-based quantitative metrics. In comparison, DSCA had a significantly lower  $\kappa$  score of 0.62, lower than all pair-wise interrater agreement scores, confirming our hypothesis that normalized measures may be more effective as a quantitative metric for not only diagnosing stenosis but also classifying it into more granular grades.

Quantitative metrics based on ratios are also inherently robust to measurement, a finding supported by our reproducibility analysis, in which the concordance correlation coefficient between R1 and R2 was 0.9 for estimating DDRDIA. Although not explicitly quantified in this study, DDRDIA may be less prone to errors because it requires the radiologist to draw 2 lines as opposed to segmentation of the dural sac for area measurements, also making it more time-efficient.

A few other quantitative ratios have been proposed in the literature. The stenosis ratio, defined as a ratio between the cross-sectional dural sac area of the motion segment and that of the stable segment, was proposed as a promising alternative to DSCA in controlling for inherent differences in patient demographics.<sup>34,35</sup>

The Torg-Pavlov ratio, which measures the ratio between the sagittal diameter of the spinal canal and the sagittal diameter of the vertebral body, is a normalized metric that can be used to assess the presence of spinal cord compression from MR imaging.<sup>36</sup> Neither the stenosis ratio nor the Torg-Pavlov ratio has been widely adopted in clinical practice for grading stenosis. The stenosis ratio requires precise measurements of multiple regions and their respective areas, which can be time-consuming and not consistent.<sup>14</sup> The Torg-Pavlov ratio has been sparsely applied to the cervical spine, with almost no prior studies establishing its effectiveness for the lumbar spine.<sup>37</sup> Moreover, for any given level, the disc level has been reported to be the most stenotic and prone to degenerative changes compared with the vertebral body.<sup>38</sup> Hence, the disc size measured as either the anterior-posterior diameter or the cross-sectional area may be a relevant feature associated with degenerative changes.

Ratios between the dural sac and the vertebral body have been published in the literature for adults as well as children and have been used for evaluation of multiple conditions.<sup>39,40</sup> For degenerative lumbar stenosis, the disc levels are the predominant stenotic



**Table 2: Classification of each spinal level as normal, mild, moderate, and severe stenosis on the evaluation cohort (n = 130) using decision trees trained on the development cohort (n = 130)**

Metric	Accuracy		AUROC		Cohen κ	
	Accuracy	95% CI	AUROC	95% CI	κ	95% CI
DSCA	64.9	(60.9–69.0)	76.6	(73.9–79.3)	0.62	(0.57–0.66)
DSDIA	71.4	(67.1–75.7)	80.9	(78.0–83.8)	0.69	(0.64–0.75)
DDRDIA	76.5	(72.6–80.4) <sup>a</sup>	84.3 <sup>a</sup>	(81.7–86.9) <sup>a</sup>	0.75 <sup>a</sup>	(0.71–0.79) <sup>a</sup>
DDRCA	78.9	(75.0–82.9) <sup>a</sup>	86.0 <sup>a</sup>	(83.3–88.5) <sup>a</sup>	0.80 <sup>a</sup>	(0.75–0.83) <sup>a</sup>

<sup>a</sup>Ratio-based metrics with higher κ scores (*P* < .001).

**Table 3: AUROC for binary classification based on each metric across symptomatic splits (VAS <7 versus VAS ≥7) of low back pain and radicular pain<sup>a</sup>**

Metric	Symptomatology Analysis of Low Back Pain				Significance <sup>a</sup> <i>P</i> Value
	VAS < 7		VAS ≥ 7		
	AUROC	95% CI	AUROC	95% CI	
Low Back Pain					
DSCA	97.7	(96.7–98.6)	95.1	(92.9–97.3)	<i>P</i> < .001
DSDIA	96.6	(95.4–97.8)	96.5	(94.9–98.1)	<i>P</i> = .43
DDRDIA	96.8	(95.6–97.9)	97.5	(96.1–98.8)	<i>P</i> < .001
DDRCA	98.5	(97.8–99.2)	96.3	(94.8–97.9)	<i>P</i> < .001
Radicular back pain					
DSCA	98.4	(97.4–99.4)	96.3	(94.5–98.1)	<i>P</i> < .001
DSDIA	98.2	(96.8–99.6)	96.9	(95.5–98.2)	<i>P</i> < .001
DDRDIA	98.7	(97.9–99.5)	97.1	(95.6–98.5)	<i>P</i> < .001
DDRCA	99.0	(98.4–99.6)	97.1	(95.5–98.7)	<i>P</i> < .001

<sup>a</sup>The *P* values represent a comparison of AUROCs among the symptomatic splits.

**Table 4: AUROC for predicting surgery using each quantitative metric on the prognostic cohort (n = 58)**

Metric	Predicting Surgery at Each Spinal Level		
	AUROC	95% CI	Significance <sup>a</sup>
DSCA	82.4	(75.5–90.4)	<i>P</i> = 1.0
DSDIA	81.3	(73.2–89.4)	<i>P</i> < .001
DDRDIA	80.8	(73.0–89.5)	<i>P</i> < .001
DDRCA	83.5 <sup>b</sup>	(76.6–90.1) <sup>b</sup>	<i>P</i> < .001 <sup>b</sup>

<sup>a</sup>The *P* values represent comparison between each metric and AUROC obtained with the baseline metric DSCA.

<sup>b</sup>Quantitative metric with the highest AUROC.

levels, motivating our proposed metrics. Studies calculating a “disc index,” a ratio of the disc-to-canal size, have reported that larger disc indexes are associated with more continuous symptoms, and as ratios decreased with time, the symptoms also regressed.<sup>41</sup> An early description of the anterior-posterior length of disc protrusion and the percentage of the canal occupied by the disc protrusion was reported in 1997, and strong predictive effects were found between ratio measurements and patient outcomes.<sup>42</sup> Subsequent studies have also supported the use of disc ratios for predicting patient groups with favorable-versus-unfavorable surgical outcomes.<sup>43</sup> While disc dimension has been previously used in the context of lumbar disc herniation, to our knowledge, it is not commonly incorporated as a potential quantitative feature along with dural sac measurements for grading LSS.

We acknowledge the following limitations of this study: Our results are based on a single expert radiologist grader and do not incorporate consensus grading or any other form of adjudication; while consensus grades are advantageous, prior studies that relied on a single grader have shown meaningful associations.<sup>28</sup> Our

approach based on decision trees may be prone to overfitting and brittle decision boundaries, wherein a slight perturbation to the development data can lead to drastically different thresholds.<sup>28</sup> Also, there are numerous statistical and machine learning techniques that can be used to determine a decision rule for each metric. A random forest model, which is a collection of several decision trees, may be more robust; we deliberately selected a decision tree for its interpretable thresholds and decision rules. We limited the depth to 3 and the maximum number of leaves to 4 to address some of the concerns around overfitting, and we observed that the derived thresholds of 91.1 mm<sup>2</sup> for severe stenosis based on DSCA and 10.1 mm for moderate or severe stenosis based on DSDIA are in line with previously published thresholds for the dural sac cross-sectional area and diameter.<sup>14</sup>

Another potential limitation is our reliance on a single outcome measure (VAS) for symptoms and a cutoff threshold of 7 to denote severe pain; other less common measures may be very valuable and the subject of future studies. Last, a potential limitation may be the acquisition of contiguous axial MR images, our institution’s routine lumbar spine imaging protocol. A prior study reported that the use of disc space-targeted angled images resulted in a 75% reduction in the detection of migrated or sequestered disc material and a 50% decrease in detected pars defects compared with contiguous axial images.<sup>44</sup>

## CONCLUSIONS

We found favorable results for our proposed ratio-based metrics, DDRDIA and DDRCA, which rely on simple measurements of the intervertebral disc and the dural sac, compared with common metrics such as the DSCA. Our results indicate that ratio-based metrics may offer a convenient trade-off between the classification



**Table 5: Classification of each spinal level as normal, mild, moderate, and severe stenosis on the prognostic cohort (n = 58) using decision trees trained on the development cohort (n = 130)**

Metric	$\kappa$ Scores for Grading Stenosis across Surgical vs Nonsurgical Levels				Significance <sup>a</sup>
	Nonsurgical Levels		Surgical Levels		
	$\kappa$	95% CI	$\kappa$	95% CI	P Value
DSCA	0.65	(0.57–0.75)	0.74	(0.63–0.85)	$P < .001$
DSDIA	0.67	(0.58–0.77)	0.68	(0.51–0.85)	$P = .23$
DDRDA	0.69	(0.61–0.76) <sup>b</sup>	0.77 <sup>b</sup>	(0.65–0.90) <sup>b</sup>	$P < .001^b$
DDRCA	0.71	(0.62–0.79)	0.73	(0.58–0.87)	$P < .001$

<sup>a</sup> The P values represent a comparison of  $\kappa$  scores between nonsurgical and surgical levels.

<sup>b</sup> The metric with the highest  $\kappa$  score for surgical levels.

of stenosis, robustness to measurement errors, and normalization across anatomic and demographic variability and stronger associations with LSS symptoms and prognosis. The proposed metrics are also practical in a clinical setting and amenable to automated estimation and can influence the diagnosis and subsequent management of patients with LSS.

## ACKNOWLEDGMENTS

We also thank Mirandarae Christine, Steven Li, and Eduarda Vieira for their contributions to image annotations.

Disclosure forms provided by the authors are available with the full text and PDF of this article at [www.ajnr.org](http://www.ajnr.org).

## REFERENCES

- Deyo RA, Gray D, Kreuter W, et al. United States trends in lumbar fusion surgery for degenerative conditions. *Spine (Phila Pa 1976)* 2005;30:1441–45 [CrossRef Medline](#)
- Cowley P. Neuroimaging of spinal canal stenosis. *Magn Reson Imaging Clin N Am* 2016;24:523–29 [CrossRef Medline](#)
- Arnoldi CC, Brodsky AE, Cauchoix J, et al. Lumbar spinal stenosis and nerve root entrapment syndromes: definition and classification. *Clin Orthop Rel Res* 1976;115:4–5 [Medline](#)
- 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;40:1033–39 [CrossRef Medline](#)
- Schroeder GD, Kurd MF, Vaccaro AR. Lumbar spinal stenosis: how is it classified? *J Am Acad Orthop Surg* 2016;24:843–52 [CrossRef Medline](#)
- Lurie J, Tomkins-Lane C. Management of lumbar spinal stenosis. *BMJ* 2016;352:h6234 [CrossRef Medline](#)
- Katz JN, Zimmerman ZE, Mass H, et al. Diagnosis and management of lumbar spinal stenosis: a review. *JAMA* 2022;327:1688–99 [CrossRef Medline](#)
- Morita M, Miyauchi A, Okuda S, et al. Comparison between MRI and myelography in lumbar spinal canal stenosis for the decision of levels of decompression surgery. *J Spinal Disord Tech* 2011;24:31–36 [CrossRef Medline](#)
- Alsaleh K, Ho D, Rosas-Arellano MP, et al. Radiographic assessment of degenerative lumbar spinal stenosis: is MRI superior to CT? *Eur Spine J* 2017;26:362–67 [CrossRef Medline](#)
- Arana E, Royuela A, Kovacs FM, et al. Lumbar spine: agreement in the interpretation of 1.5-T MR images by using the Nordic Modic Consensus Group classification form. *Radiology* 2010;254:809–17 [CrossRef Medline](#)
- Schizas C, Theumann N, Burn A, et al. Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images. *Spine (Phila Pa 1976)* 2010;35:1919–24 [CrossRef Medline](#)
- Khalsa SS, Kim HS, Singh R, et al. Radiographic outcomes of endoscopic decompression for lumbar spinal stenosis. *Neurosurg Focus* 2019;46:E10 [CrossRef Medline](#)
- Andreisek G, Imhof M, Wertli M, et al; Lumbar Spinal Stenosis Outcome Study Working Group Zurich. A systematic review of semi-quantitative and qualitative radiologic criteria for the diagnosis of lumbar spinal stenosis. *AJR Am J Roentgenol* 2013;201:W735–46 [CrossRef Medline](#)
- Steurer J, Roner S, Gnannt R, et al. Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: a systematic literature review. *BMC Musculoskelet Disord* 2011;12:175 [CrossRef Medline](#)
- Fukusaki M, Kobayashi I, Hara T, et al. Symptoms of spinal stenosis do not improve after epidural steroid injection. *Clin J Pain* 1998;14:148–51 [CrossRef Medline](#)
- Koc Z, Ozcakar S, Sivrioglu K, et al. Effectiveness of physical therapy and epidural steroid injections in lumbar spinal stenosis. *Spine (Phila Pa 1976)* 2009;34:985–89 [CrossRef Medline](#)
- Hamanishi C, Matukura N, Fujita M, et al. Cross-sectional area of the stenotic lumbar dural tube measured from the transverse views of magnetic resonance imaging. *J Spinal Disord* 1994;7:388–93 [Medline](#)
- Mariconda M, Fava R, Gatto A, et al. Unilateral laminectomy for bilateral decompression of lumbar spinal stenosis: a prospective comparative study with conservatively treated patients. *J Spinal Disord Tech* 2002;15:39–46 [CrossRef Medline](#)
- Jönsson B, Annertz M, Sjöberg C, et al. A prospective and consecutive study of surgically treated lumbar spinal stenosis, Part I: clinical features related to radiographic findings. *Spine* 1997;22:2932–37 [CrossRef Medline](#)
- Dora C, Wälchli B, Elfering A, et al. The significance of spinal canal dimensions in discriminating symptomatic from asymptomatic disc herniations. *Eur Spine J* 2002;11:575–81 [CrossRef Medline](#)
- Aalto TJ, Malmivaara A, Kovacs F, et al. Preoperative predictors for postoperative clinical outcome in lumbar spinal stenosis: systematic review. *Spine (Phila Pa 1976)* 2006;31:E648–63 [CrossRef Medline](#)
- Andreisek G, Hodler J, Steurer J. Uncertainties in the diagnosis of lumbar spinal stenosis. *Radiology* 2011;261:681–84 [CrossRef Medline](#)
- Friedly JL, Jarvik JG. Agreeing (or not) on how to describe spinal stenosis: expanding a narrow mindset. *Radiology* 2012;264:3–4 [CrossRef Medline](#)
- Miskin N, Gaviola GC, Huang RY, et al. Intra- and intersubspecialty variability in lumbar spine MRI interpretation: a multireader study comparing musculoskeletal radiologists and neuroradiologists. *Curr Probl Diagn Radiol* 2020;49:182–87 [CrossRef Medline](#)
- Mamisch N, Brumann M, Hodler J, et al. Radiologic criteria for the diagnosis of spinal stenosis: results of a Delphi survey. *Radiology* 2012;264:174–79 [CrossRef Medline](#)
- Downie WW, Leatham PA, Rhind VM, et al. Studies with pain rating scales. *Ann Rheum Dis* 1978;37:378–81 [CrossRef Medline](#)
- Breiman L, Friedman JH, Olshen RA, et al. *Classification and Regression Trees*. Routledge; 1984:368
- Huber FA, Stutz S, Martini IVd, et al. Qualitative versus quantitative lumbar spinal stenosis grading by machine learning supported texture analysis: experience from the LSOS study cohort. *Eur J Radiol* 2019;114:45–50 [CrossRef Medline](#)
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825–30

30. Virtanen P, Gommers R, Oliphant TE, et al; SciPy 1.0 Contributors. **SciPy 1.0: fundamental algorithms for scientific computing in Python.** *Nat Methods* 2020;17:261–72 [CrossRef Medline](#)
31. DeLong ER, DeLong DM, Clarke-Pearson DL. **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988;44:837–45 [CrossRef Medline](#)
32. Buser Z, Ortega B, D'Oro A, et al. **Spine degenerative conditions and their treatments: national trends in the United States of America.** *Global Spine J* 2018;8:57–67 [CrossRef Medline](#)
33. Nuttall FQ. **Body mass index obesity, BMI, and health: a critical review.** *Nutr Today* 2015;50:117–28 [CrossRef Medline](#)
34. Laurencin CT, Lipson SJ, Senatus P, et al. **The stenosis ratio: a new tool for the diagnosis of degenerative spinal stenosis.** *Int J Surg Investig* 1999;1:127–31 [Medline](#)
35. Kitab S, Lee BS, Benzel EC. **Redefining lumbar spinal stenosis as a developmental syndrome: an MRI-based multivariate analysis of findings in 709 patients throughout the 16- to 82-year age spectrum.** *J Neurosurg Spine* 2018;29:654–60 [CrossRef Medline](#)
36. Pavlov H, Torg JS, Robie B, et al. **Cervical spinal stenosis: determination with vertebral body ratio method.** *Radiology* 1987;164:771–75 [CrossRef Medline](#)
37. Bajwa NS, Toy JO, Ahn NU. **Application of a correlation between the lumbar Torg ratio and the area of the spinal canal to predict lumbar stenosis: a study of 420 postmortem subjects.** *J Orthopaed Traumatol* 2013;14:207–12 [CrossRef Medline](#)
38. Thomé C, Börm W, Meyer F. **Degenerative lumbar spinal stenosis: current strategies in diagnosis and treatment.** *Dtsch Arztebl Int* 2008;105:373–79 [CrossRef Medline](#)
39. Knirsch W, Kurtz C, Häffner N, et al. **Normal values of the sagittal diameter of the lumbar spine (vertebral body and dural sac) in children measured by MRI.** *Pediatr Radiol* 2005;35:419–24 [CrossRef Medline](#)
40. Pierro A, Cilla S, Maselli G, et al. **Sagittal normal limits of lumbosacral spine in a large adult population: a quantitative magnetic resonance imaging analysis.** *J Clin Imaging Sci* 2017;7:35 [CrossRef Medline](#)
41. Fagerlund MK, Thelander U, Friberg S. **Size of lumbar disc hernias measured using computed tomography and related to sciatic symptoms.** *Acta Radiol* 1990;31:555–58 [CrossRef Medline](#)
42. Carragee EJ, Kim DH. **A prospective analysis of magnetic resonance imaging findings in patients with sciatica and lumbar disc herniation: correlation of outcomes with disc fragment and canal morphology.** *Spine* 1997;22:1650–60 [CrossRef Medline](#)
43. Varlotta CG, Manning JH, Ayres EW, et al. **Preoperative MRI predictors of health-related quality of life improvement after microscopic lumbar discectomy.** *Spine J* 2020;20:391–98 [CrossRef Medline](#)
44. Singh K, Helms CA, Fiorella D, et al. **Disc space-targeted angled axial MR images of the lumbar spine: a potential source of diagnostic error.** *Skeletal Radiol* 2007;36:1147–53 [CrossRef Medline](#)