European Spine Journal (2025) 34:869–873
https://doi.org/10.1007/s00586-024-08612-z

RESEARCH

Inter-observer variability in the classification of lumbar foraminal stenosis in magnetic resonance imaging using different evaluation scales

José Sá Silva[1] · Ana Pereira[1] · Vasco Abreu[1] · João Pedro Filipe[1]

Received: 3 October 2024 / Revised: 3 October 2024 / Accepted: 11 December 2024 / Published online: 20 December 2024

Abstract

Background The evaluation of lumbar spine degeneration on magnetic resonance imaging (MRI) is prone to inter-reader variability, including when assessing foraminal changes. This variability, often due to subjective criteria and inconsistent terminology, may affect clinical correlations. Standardized criteria could help improve agreement among readers.

Materials and methods MRI of the lumbar spine of 50 randomly selected patients were evaluated by 12 independent readers. Foraminal stenosis was assessed using four different rating scales for each patient. The first scale classified stenosis as presence/absence of neurologic compromise of the spinal nerve root at the foramen, the second scale classified stenosis as absent/mild/moderate/severe, the third scale as normal/contact of disk or osteophyte with the nerve root/deviation of the nerve root/compression of the nerve root, and the fourth scale utilized the Lee et al. criteria. Agreement analysis was performed using Fleiss' kappa coefficients.

Results Agreement was moderate using the first scale ($k = 0.439$), and significantly lower using the second, third and fourth scales ($k = 0.310$, $k = 0.311$, $k = 0.295$, respectively). When comparing the agreements obtained between board certified neuroradiologists and between neuroradiology residents, there was statistically significant differences when using the third and fourth scales, where the agreement for board certified neuroradiologists was higher, but still only fair. Individual kappas showed that in the second, third, and fourth scales the levels of agreement were higher in the extremes of the scale, namely, when there was no stenosis or when the stenosis was maximal with nerve compression.

Conclusions Levels of agreement can differ depending on the scale used. Simpler dichotomous scales may return higher levels of agreement compared to more complex ones. For the non-dichotomous scales, using different scales may not result in overall different levels of agreement. Given the overall low inter-rater agreements observed, there is probably significant potential to enhance agreement through more rigorous training and consensus-building.

Keywords Inter-observer variability · Inter-observer agreement · Rating scale · Lumbar spine · Foraminal stenosis · Spine degenerative disease · Magnetic resonance

Introduction

The evaluation of degenerative spine disease on magnetic resonance imaging (MRI) is subject to inter-reader variability, and the assessment of foraminal degenerative changes, although less studied in the available literature, is also prone to inter-reader variation [1–3]. This can have clinical implications because the degenerative abnormalities that have less inter-rater agreement might also show lower correlation with clinical symptoms [4].

Different readers might use different criteria when examining degenerative spine disease, and these criteria might

✉ José Sá Silva
jose.msas.silva@gmail.com

Ana Pereira
gemeasanaisabel@gmail.com

Vasco Abreu
vrlsa1993@gmail.com

João Pedro Filipe
mailjpfilipe@gmail.com

[1] Department of Neuroradiology, Centro Hospitalar Universitário de Santo António, Unidade Local de Saúde de Santo António, Porto, Portugal

even be inconsistently applied because many of them are based on subjective evaluation. Different terminology use is also a problem, and even though several degenerative disc disease classifications tried to resolve this problem, there is still a lack of consensus for foraminal stenosis nomenclature [5–8].

The use of standardized imaging criteria might improve rapport between readers and referring physicians and reduce the variability of subjective interpretation of imaging findings. Thus, it is important to understand if the use of clearer definitions and standardized criteria does improve inter-reader agreement. For that purpose, the goal of our study was to evaluate and compare the inter-reader variability of foraminal stenosis classification in lumbar spine MRI using different classification criteria, ranging from a simpler and more open classification to more complex and restrictive ones.

Materials and methods

Patient sample

This study utilizes retrospective data, with institutional ethical committee approval and waiver of informed consent. A total of 50 MRI scans of the lumbar spine done in the previous year were randomly selected from our institutional Picture Archiving and Communication System. Inclusion criteria were age above 18 and the examination being performed due to clinical suspicion of symptoms associated with degenerative spine disease. Postoperative and patients with implanted hardware were excluded.

MR imaging and assessment criteria

All examinations were performed on one of two different MRI scanners with field strengths at 1.5 and 3T, manufactured by GE (Milwaukee, WI, USA) and Philips (Amsterdam, The Netherlands), respectively. All studies included sagittal and axial T1 and T2 weighted images.

For each of the 50 patients, only one intervertebral foramen was evaluated, which was randomly chosen between the four foramina of the L4-L5 and L5-S1 levels, as these levels are the most commonly affected by degenerative changes.

At the beginning of the image analysis, readers were provided with the descriptions of the evaluation criteria they were to use. Four different scales were used in each patient, and comprised of the following sets of criteria: (1) the reader was asked if he thought there was (or not) neurologic compromise of the spinal nerve root at the intervertebral foramen; (2) the reader was asked to classify the foraminal stenosis as absent, mild, moderate or severe; (3) the reader was asked to apply an adapted form of the Pfirrmann et al. classification [7] in which they classified the foraminal stenosis as normal, contact of disk material/osteophyte with the spinal nerve root, deviation of the spinal nerve root, or compression of the spinal nerve root; (4) finally, the reader was asked to apply the Lee et al. classification [9], which classifies lumbar foraminal stenosis into grades 0, 1, 2 and 3 according to perineural fat obliteration and morphologic changes of the nerve root.

The Pfirrmann et al. classification is popular among radiologists and has been previously recommended for the evaluation of foraminal nerve root impingement [10]. The Lee et al. classification was applied as originally described. This classification is a grading system developed for lumbar foraminal stenosis on the basis of sagittal MRI and is associated with high inter-reader agreement [9].

Image analysis

Image analysis was performed by 12 readers from our hospital center. Seven readers were board certified neuroradiologists (with an average of about 12 years of neuroradiology experience) and five readers were neuroradiology residents (all with experience between three and five years).

All reading sessions were performed using anonymized DICOM data. All readers were blinded to the results of other readers and to all patient data including age, sex, and clinical symptoms. Image analysis was performed only on standard sagittal T1 and T2 weighted images, and on axial T1 and T2 weighted images.

Data and statistical analysis

For statistical analysis we used SPSS Statistics (version 29; IBM Corp., Armonk, NY, USA). Each one of the four different scales for the evaluation of foraminal stenosis was assessed for inter-reader agreement using Fleiss' kappa. We first performed kappa analysis for all our readers, and then separated board certified neuroradiologists from neuroradiology residents and calculated the kappa coefficients for each group. We analyzed the individual kappas to assess the level of agreement between our readers for each of the categories of the response variable in each scale. For all analyses, non-overlapping 95% confidence intervals were considered as a statistically significant difference. According to the standard originally proposed by Landis and Koch, Fleiss' kappa values below 0 were interpreted as poor agreement, 0.01–0.20 as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement, 0.61–0.80 as substantial agreement, and 0.81-1.00 as almost perfect agreement [11].

Table 1 Fleiss' kappa coefficients for inter-reader agreement in classifying foraminal stenosis, using different scales

	Scale 1	Scale 2	Scale 3	Scale 4
All readers	**0.439** (95% CI, 0.405–0.473)	**0.310** (95% CI, 0.289–0.330)	**0.311** (95% CI 0.291–0.332)	**0.295** (95% CI 0.274–0.316)
Board certified neuroradiologists	**0.467** (95% CI 0.406–0.527)	**0.300** (95% CI 0.263–0.337)	**0.341** (95% CI 0.305–0.378)	**0.328** (95% CI 0.291–0.365)
Neuroradiology residents	**0.365** (95% CI 0.278–0.453)	**0.314** (95% CI 0.261–0.366)	**0.250** (95% CI 0.197–0.303)	**0.225** (95% CI 0.172–0.278)

Table 2 Individual kappas for each of the categories of each scale

Scale 1		Scale 2		Scale 3		Scale 4	
Compromise of the spinal nerve root at the intervertebral foramen	0.439	Normal	0.400	Normal	0.494	Grade 0	0.442
		Mild	0.191	Contact of disk/osteophyte with the spinal nerve root	0.175	Grade 1	0.156
No compromise of the spinal nerve root	0.439	Moderate	0.157	Deviation of the spinal nerve root	0.048	Grade 2	0.530
		Severe	0.489	Compression of the spinal nerve root	0.436	Grade 3	0.394

Results

The study population consisted of 50 patients, 25 males, with age ranging from 27 to 84 years with a mean age of 62.7. A total of 50 intervertebral foramina were assessed by our 12 readers, resulting in a total of 600 classifications. Kappa coefficients for all readers, for board certified neuroradiologists, and for neuroradiology residents are summarized in Table 1. Individual kappas are reported in Table 2.

First evaluation scale

In this scale, the reader was asked if he thought there was (or not) neurologic compromise of the spinal nerve root at the intervertebral foramen. For all readers, the agreement was k = 0.439 (95% CI, 0.405–0.473), $p < 0.001$. For board certified neuroradiologists the agreement was k = 0.467 (95% CI 0.406–0.527), $p < 0.001$. For neuroradiology residents the agreement was k = 0.365 (95% CI 0.278–0.453), $p < 0.001$.

Second evaluation scale

With this scale, the reader classified the foraminal stenosis as absent, mild, moderate or severe. For all readers, the agreement was k = 0.310 (95% CI, 0.289–0.330), $p < 0.001$. For board certified neuroradiologists the agreement was k = 0.300 (95% CI 0.263–0.337), $p < 0.001$. For neuroradiology residents the agreement was k = 0.314 (95% CI 0.261–0.366), $p < 0.001$. Individual kappas for "normal", "mild", "moderate", and "severe" were 0.400, 0.191, 0.157, and 0.489, respectively.

Third evaluation scale

In this scale, the reader was asked to classify the foraminal stenosis as normal, contact of disk material/osteophyte with the spinal nerve root, deviation of the spinal nerve root, or compression of the spinal nerve root. For all readers, the agreement was k = 0.311 (95% CI 0.291–0.332), $p < 0.001$. For board certified neuroradiologists the agreement was k = 0.341 (95% CI 0.305–0.378), $p < 0.001$. For neuroradiology residents the agreement was k = 0.250 (95% CI 0.197–0.303), $p < 0.001$. Individual kappas for "normal", "contact of disk material/osteophyte with the spinal nerve root", "deviation of the spinal nerve root", and "compression of the spinal nerve root" were 0.494, 0.175, 0.048, and 0.436, respectively.

Fourth evaluation scale

This scale asked the reader to apply the Lee et al. classification, which divides lumbar foraminal stenosis into grades 0, 1, 2 and 3 [9]. For all readers, the agreement was k = 0.295 (95% CI 0.274–0.316), $p < 0.001$. For board certified neuroradiologists the agreement was k = 0.328 (95% CI 0.291–0.365), $p < 0.001$. For neuroradiology residents the agreement was k = 0.225 (95% CI 0.172–0.278), $p < 0.001$. Individual kappas for grade 0, grade 1, grade 2, and grade 3 were 0.442, 0.156, 0.053, and 0.394, respectively.

Discussion

Our results demonstrate that levels of agreement can differ depending on the scale used. When using the first scale, the agreement between all readers was moderate and was the highest compared to the other scales. For the second, third, and fourth scales, the agreement was only fair and was significantly lower than the agreement observed in the first scale. The first scale was dichotomous and was the simpler to apply. For the non-dichotomous and more complex scales, using different scales for classifying lumbar

foraminal stenosis did not lead to different overall agreement. This finding aligns with the concept that simpler classification methods may be easier to reproduce [12], and that ambiguous scales may return biased classifications by the raters [13].

When separately analyzing the agreement obtained by the board certified neuroradiologists and by the neuroradiology residents, these groups didn't obtain significantly different results from the group of all readers combined. However, when comparing the agreement obtained between board certified neuroradiologists with the agreement obtained between neuroradiology residents, there was statistically significant differences when using the third and fourth scales, where the agreement for board certified neuroradiologists was higher. The third and fourth scales are more complex, and board-certified neuroradiologists, due to their greater experience, may have applied the criteria more effectively, resulting in better reproducibility.

The analysis of the individual kappas showed that in all non-dichotomous scales the levels of agreement were higher in the extremes of the scale, namely, when there was no stenosis or when the stenosis was maximal with nerve compression. Extreme response tendencies on item scales independent of item content have been reported, and the use of bipolar scales can influence those tendencies [14–16]. If not considered when constructing classification scales, these factors may bias their utilization.

Several other studies evaluated the variability of reporting degenerative findings of the spine. It should be noted that comparison with other studies is in some instances hampered by the use of different statistical methods to assess variability. Miskin et al. evaluated the agreement of lumbar foraminal stenosis assessment using a 5-point ordinal scale, obtaining a maximum agreement of 0.670 (95% CI, 0.625 to 0.714) between neuroradiologists, considered moderate, using Cohen's kappa [2]. Fu et al. tested the agreement between multiple raters in assessing lumbar foraminal stenosis using a standardized ordinal scale with three categories, obtaining a moderate agreement of 0.481 (95% CI, 0.472 to 0.490), using Fleiss's kappa [1]. Lurie et al. evaluated lumbar foraminal stenosis using 4-item ordinal scales similar to the ones we used. In their study, foraminal stenosis was rated as "none," "mild," "moderate," or "severe", and the degree of nerve root impingement was rated as "none," "touching," "displacing," or "compressing". Using weighted kappa statistics, foraminal stenosis showed an overall moderate agreement of 0.58 (95% CI, 0.53 to 0.63), and nerve root impingement showed an overall moderate agreement of 0.51 (95% CI, 0.42 to 0.59) [17]. Winklhofer et al. investigated foraminal nerve root impingement using a scale based on the Pfirrmann et al. classification [7], similar to the one we used. They obtained values between 0.59 and 0.67 for the inter-reader agreement using Cohen's kappa [3]. All these studies utilized the standard proposed by Landis and Koch to interpret the degree of agreement [11].

When using non-dichotomous scales, the readers in our study obtained only fair agreement, showing an overall lower agreement compared to the aforementioned reports. The cause of this difference is hard to pinpoint, but may reflect some lack of reproducibility between our participants in objectifying the assessment of foraminal stenosis. Considering this, another study by Miskin et al. demonstrated that when the raters were trained on a standardized classification of degenerative change on MRI, inter-rater agreement increased [18]. The lower agreement values in our study suggest that training provided previously to the application of classification scales, which didn't occur in our sample, may be an important factor in reducing subjectivity and, consequently, variability.

The study executed by Lee et al., whose grading system for lumbar foraminal stenosis we used as our fourth classification scale, obtained almost perfect agreement between raters [9]. This different result may be due to an incorrect application of the grading system by the raters in our study. However, this once again suggests that the application of grading scales by different sets of raters may be influenced by subjectivity.

One of the main limitations of our study is the absence of symptomatic correlation and surgical information. Studying the relation between our ratings and the clinical symptoms could show if clinically relevant abnormalities on imaging are being reported as such. The lack of an accepted gold-standard for reporting lumbar spine degenerative disease on MRI can also represent a limitation to test grading scales of degenerative abnormalities. Another limitation is that the numbers of years of experience varied importantly between our raters, and although we did a separate analysis for board certified neuroradiologists and neuroradiology residents, the differences in experience may have contributed to variability. Generalization of our results is limited, since our raters represent a single center small sample that may be influenced by unrecognized local biases.

Conclusions

The level of agreement in the evaluation of foraminal stenosis on MRI using rating scales was fair to moderate. Levels of agreement can differ depending on the scale used. Simpler dichotomous scales may return higher levels of agreement compared to more complex ones. For the non-dichotomous scales, using different scales may not result in overall different levels of agreement, but board certified neuroradiologists show higher levels of agreement using more complex

scales when compared to neuroradiology residents. In all non-dichotomous scales, the level of agreement tends to be higher in the extremes of the scale, namely, when there is no stenosis or when the stenosis is maximal with nerve compression.

Given the overall low inter-rater agreements observed, there is probably still significant potential to enhance agreement through more rigorous training and consensus-building.

Author contributions JSS: Conception and writing of the manuscript, data curation, methodology.AP: Conceptualization, data curation.VA: Resources, validation, review.JPF: Data interpretation, resources, supervision, validation, review and editing.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Fu MC et al (2014) Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. Spine J 14(10):2442–2448
2. Miskin N et al (2020) Intra- and Intersubspecialty Variability in Lumbar Spine MRI Interpretation: A Multireader Study Comparing Musculoskeletal Radiologists and Neuroradiologists. Curr Probl Diagn Radiol 49(3):182–187
3. Winklhofer S et al (2017) Degenerative lumbar spinal canal stenosis: intra- and inter-reader agreement for magnetic resonance imaging parameters. Eur Spine J 26(2):353–361
4. Moojen WA et al (2018) Preoperative MRI in Patients With Intermittent Neurogenic Claudication: Relevance for Diagnosis and Prognosis. Spine (Phila Pa 1976) 43(5):348–355
5. Arana E et al (2010) Lumbar spine: agreement in the interpretation of 1.5-T MR images by using the Nordic Modic Consensus Group classification form. Radiology 254(3):809–817
6. Fardon DF et al (2014) Lumbar disc nomenclature: version 2.0: Recommendations of the combined task forces of the North American Spine Society, the American Society of Spine Radiology and the American Society of Neuroradiology. Spine J 14(11):2525–2545
7. Pfirrmann CW et al (2004) MR image-based grading of lumbar nerve root compromise due to disk herniation: reliability study with surgical correlation. Radiology 230(2):583–588
8. Ross JS (2010) Babel 2.0. Radiology 254(3):640–641
9. Lee S et al (2010) A practical MRI grading system for lumbar foraminal stenosis. AJR Am J Roentgenol 194(4):1095–1098
10. Andreisek G et al (2014) Consensus conference on core radiological parameters to describe lumbar stenosis - an initiative for structured reporting. Eur Radiol, 24(12): pp. 3224-32
11. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
12. Gwet K (2012) Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters
13. Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cogn Psychol 5(3):213–236
14. Van Vaerenbergh Y, Thomas TD (2012) Response Styles in survey research: A literature review of antecedents, consequences, and remedies. Int J Public Opin Res 25(2):195–217
15. Greenleaf ea (1992) measuring extreme response style. Pub Opin Q 56(3):328–351
16. Lau MY-K (2008) Extreme response style: An empirical investigation of the effects of scale response format and fatigue. University of Notre Dame
17. Lurie JD et al (2008) Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine (Phila Pa 1976) 33(14):1605–1610
18. Miskin N et al (2022) Standardized classification of lumbar spine degeneration on magnetic resonance imaging reduces intra- and inter-subspecialty variability. Curr Probl Diagn Radiol 51(4):491–496