ARTICLE IN PRESS

Current Problems in Diagnostic Radiology 000 (2021) 1–6

Current Problems in Diagnostic Radiology

journal homepage: www.cpdrjournal.com

Standardized Classification of Lumbar Spine Degeneration on Magnetic Resonance Imaging Reduces Intra- and Inter-subspecialty Variability

Nityanand Miskin, MD*[,1], Glenn C. Gaviola, MD, Raymond Y. Huang, MD, PhD, Christine J. Kim, MD, Thomas C. Lee, MD, Kirstin M. Small, MD, Ged G. Wieschhoff, MD, Jacob C. Mandell, MD

Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

A B S T R A C T

Background and Purpose: To determine the efficacy of standardized definitions of degenerative change in reducing variability in interpretation of lumbar spine magnetic resonance imaging within and between groups of subspecialty-trained neuroradiologists (NR) and musculoskeletal radiologists (MSK).
Materials and Methods: Six radiologists, three from both NR and MSK groups were trained on a standardized classification system of degenerative change. After an 11-month washout period, they independently re-interpreted fifty exams at the L4-L5 and L5-S1 levels. Responses were converted to a six-point ordinal scale for the assessment of neural foraminal stenosis and spinal canal stenosis (SCS), three-point scale for lateral recess stenosis, and four-point scale for facet osteoarthritis (FO). Intra-subspecialty and inter-subspecialty analysis was performed using the weighted Cohen's kappa with a binary matrix of all reader pairs.
Results: Inter-subspecialty agreement improved from k=0.527 (moderate) to k=0.602 (substantial) for neural foraminal stenosis, from k=0.540 (moderate) to k=0.652 (substantial) for SCS, from k=0.0818 (slight) to k=0.337 (fair) for lateral recess stenosis, and from k=0.176 (slight) to k=0.495 (moderate) for FO. The NR group demonstrated improved intra-subspecialty agreement for the assessment of SCS, from k=0.368 (fair) to k=0.638 (substantial). The MSK group demonstrated improved intra-subspecialty agreement for the assessment of FO, from k=0.134 (slight) to k=0.413 (moderate). Intra-subspecialty agreement was similar for other parameters before and after training.
Conclusions: As result of the standardized definitions training, the NR and MSK groups each improved in one of the four parameters, while inter-subspecialty variability improved in all four parameters. These definitions may be useful in clinical practice across radiology subspecialties.

Introduction

Magnetic resonance imaging (MRI) of the lumbar spine is the recommended imaging study of choice for treatment planning in patients with chronic low back pain, as well as patients with lumbar radiculopathy and signs of spinal canal stenosis.[1,2] Accurate characterization of lumbar spine MRI findings is crucial for adequate pre-procedural evaluation of patients for potential invasive spine interventions, including lumbar decompressive surgery and injections.[1-4] Unfortunately, reliable and accurate descriptions determining the severity of lumbar spine degenerative changes is challenging.

Prior work has demonstrated substantial variability in lumbar spine interpretation, both within the same rater and between raters.[5-8] Interrater variability has been demonstrated between radiologists,[8-10] radiologists and spine clinicians,[4,11] and between spine clinicians.[5,11] Importantly, variability in the reporting of lumbar spine degenerative change degrades its utility in correlating findings with clinical symptoms.[12]

Compounding this variability within radiology is the workflow present even in highly subspecialized institutions, where lumbar spine MRI is read by both neuroradiology and musculoskeletal radiology divisions. We recently reported high variability between musculoskeletal (MSK) radiologists and neuroradiologists (NR) groups in lumbar spine MRI interpretation, which we defined as "inter-subspecialty variability".[13] Furthermore, we demonstrated extensive variability *within* the MSK and NR groups, defined as "intra-subspecialty variability." The identification of the presence and degree of this variability demands a much-needed solution to help minimize reader discrepancies.

We recently published a classification system consisting of the most clinically meaningful parameters described in lumbar spine degeneration, as determined by multidisciplinary consensus of a group made up of radiologists and non-radiology spine experts.[14] One such solution to the problem of intra- and inter-subspecialty variability in lumbar spine MRI interpretation is the use of this standardized classification system. Because the classification system was formulated with direct input from spine clinicians to highlight the most clinically important parameters, a uniform

Abbreviations: MRI, magnetic resonance imaging; NR, neuroradiology; MSK, musculoskeletal radiology; SCS, spinal canal stenosis; NFS, neural foraminal stenosis; LRS, lateral recess stenosis; FO, facet joint osteoarthritis

The author names are written as they would like their name to appear.
This study was not funded by any source.
The authors declare no conflicts of interest.

*Reprint requests: Nityanand Miskin, MD, Department of Radiology, Brigham and Women's Hospital, 75 Francis St, Boston, MA 02115.
E-mail address: nmiskin@bwh.harvard.edu (N. Miskin).
[1]@NandMiskinMD

https://doi.org/10.1067/j.cpradiol.2021.08.001
0363-0188/© 2021 Elsevier Inc. All rights reserved.

means to characterize lumbar spine degeneration could facilitate concordant descriptions of relevant findings within radiology divisions and in communications with ordering clinicians.

The purpose of this study was to assess the efficacy of a standardized classification in reducing both intra-subspecialty and inter-subspecialty variability in the reporting of lumbar spine degeneration.

Materials and Methods

This study was HIPAA-compliant, used retrospective data, and received Institutional Review Board approval, with a waiver of informed consent. The same 50 consecutive patient examinations utilized in the prior study were used in this study.[13] These patients underwent an MRI of the lumbar spine performed without the use of intravenous gadolinium contrast material, and were performed at our institution for the indication of low back pain lasting longer than 6 weeks, radiculopathy, or clinical symptoms of spinal stenosis. Examinations that were performed with the indication of primary tumor or metastatic disease, suspected infection, and neurological disorders were excluded. Patients with a history of hardware implantation were also excluded.

MRI Exams

All MRI examinations were performed on one of 10 clinical MRI scanners (five at 1.5 T and five at 3.0 T), manufactured by GE (Milwaukee, WI) or Siemens (Erlangen, Germany). Forty of the exams were performed at 3.0 T and 10 were performed at 1.5 T. Both magnet strengths were included given both groups interpret examinations at these magnet strengths as part of daily clinical practice, as do many radiologists in general who interpret MRI studies. Each non-contrast lumbar spine MRI included the following sequences: localizer, axial T1-weighted, axial T2-weighted, sagittal T1-weighted, sagittal T2-weighted, sagittal STIR.

Lumbar Spine MRI Reporting

The same six radiologists who participated in the prior study served as readers for this study.[13] This consisted of three fellowship-trained neuroradiologists and three fellowship-trained MSK radiologists, who reviewed all 50 MRI studies in an independent manner. The NR group consisted of a reader with 4 years of attending radiology experience, a reader with 8 years of attending radiology experience, and a reader with 11 years of attending radiology experience. The MSK group consisted of a reader with 2 years of attending radiology experience, a reader with 10 years of attending radiology experience, and another reader with 10 years of attending radiology experience. As in the prior study, all readers in this study except for 1 radiologist reader in the NR group received their fellowship training at our institution. As in the prior study, the reading sessions were performed utilizing deidentified DICOM images on a secure web-based Picture Archiving and Communication System (PACS) integrating data collection with data display.[15] Given their participation in the prior study, all six readers were familiar with using the web-based PACS interface necessary to interpret the L4-L5 and L5-S1 disc levels for each patient's scan. To replicate the usual clinical workflow of study interpretation, all six readers were again instructed to interpret the findings at the L4-L5 and L5-S1 disc levels for each scan exactly as the readers would in clinical practice. Determination of the L5-S1 level was left to the discretion of the interpreting reader, to best replicate the clinical environment encountered in real-life clinical practice. The L4-5 level, as a result, was determined in relation to the assigned L5-S1 level. The order of the 50 patient scans were randomized on the web-based PACS interface. The readers underwent a washout period of 11 months from the prior study before re-interpreting the images. All six readers were blinded to all identifying patient data, including age, sex, laterality, and suspected level of clinical symptoms.

Standardized Definitions

Four parameters of lumbar spine degeneration were assessed, including spinal canal stenosis (SCS), neural foraminal stenosis (NFS), lateral recess stenosis (LRS), and facet osteoarthritis (FO). These parameters were intentionally selected given their clinical relevance, as each of these locations can be therapeutically targeted by injections or decompressive surgery. The standardized definitions of four parameters were created by multidisciplinary consensus, after consultation with orthopedic and neurological spine surgeons, physiatrists, and radiologists.[14] The overview of the schematic diagrams summarizing the MRI-based classification system for all four parameters of lumbar spine degeneration is shown in Figure 1. The readers were provided with (1) schematic diagrams and representative MR images (separate from the aforementioned 50 cases), which were combined into a Microsoft PowerPoint file (Redmond, WA), (2) a single-page summary PDF (Adobe, Mountainview, CA) with schematic diagrams and descriptions, and (3) a recorded a 19-minute video presenting the classification, which can be accessed at the following link: https://www.youtube.com/watch?v=Cn6z42uo7hk.

Report Analysis

Using the same ordinal scale as in the prior study, the free-text interpretations were analyzed and converted into a six-point scale for the assessment of NFS and SCS: 0=not reported or stated as absent, 1=mild, 2=mild-to-moderate, 3=moderate, 4=moderate-to-severe, 5=severe, with the intermediate gradings of mild-to-moderate and moderate-to-severe assigned at the discretion of the readers (and with examples given in the provided video). The grading of LRS was converted to a three-point ordinal scale for each side: 0=not reported or stated as absent, 1=abutment or contact of the descending nerve root, 2=compression of the descending nerve root. The grading of right and/or left FO was converted to a four-point ordinal scale for each level: 0=not reported or stated as absent, 1=mild, 2=moderate, 3=severe. The compilation of the converted free-text was recorded in Microsoft Excel (Microsoft, Redmond, Washington).

The prior study's ordinal scale for the parameters of LRS and FO consisted of denoting its absence of presence in reporting (0 or 1), which differs in comparison to the current study.

Data Analysis

Statistical analyses were performed by using R 3.4.1 (R Foundation for Statistical Computing, Vienna, Austria). NFS, SCS, LRS, and FO were assessed for intra-subspecialty and inter-subspecialty agreement by using the weighted Cohen's kappa coefficient. For intra-subspecialty (IS) agreement, this was performed by calculating a single value with 95% confidence intervals for a matrix consisting of the three paired responses (A-B, A-C, and B-C pairs concatenated into two columns consisting of AAB and BCC). Inter-subspecialty agreement was assessed also by calculating a single kappa coefficient with 95% confidence intervals for the nine unique NR-MSK pairs of readers. The Cohen's kappa coefficient was interpreted using the scale established by Landis and Koch[16]: kappa < 0 is poor agreement; 0.00-0.20 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60, moderate agreement; 0.61-0.80, substantial agreement; and 0.81-1.00, almost perfect agreement.

Results

Neural Foraminal Stenosis

Two hundred neural foramina (two sides at two levels for the 50 exams) were assessed by each of the six readers for a total of 1200

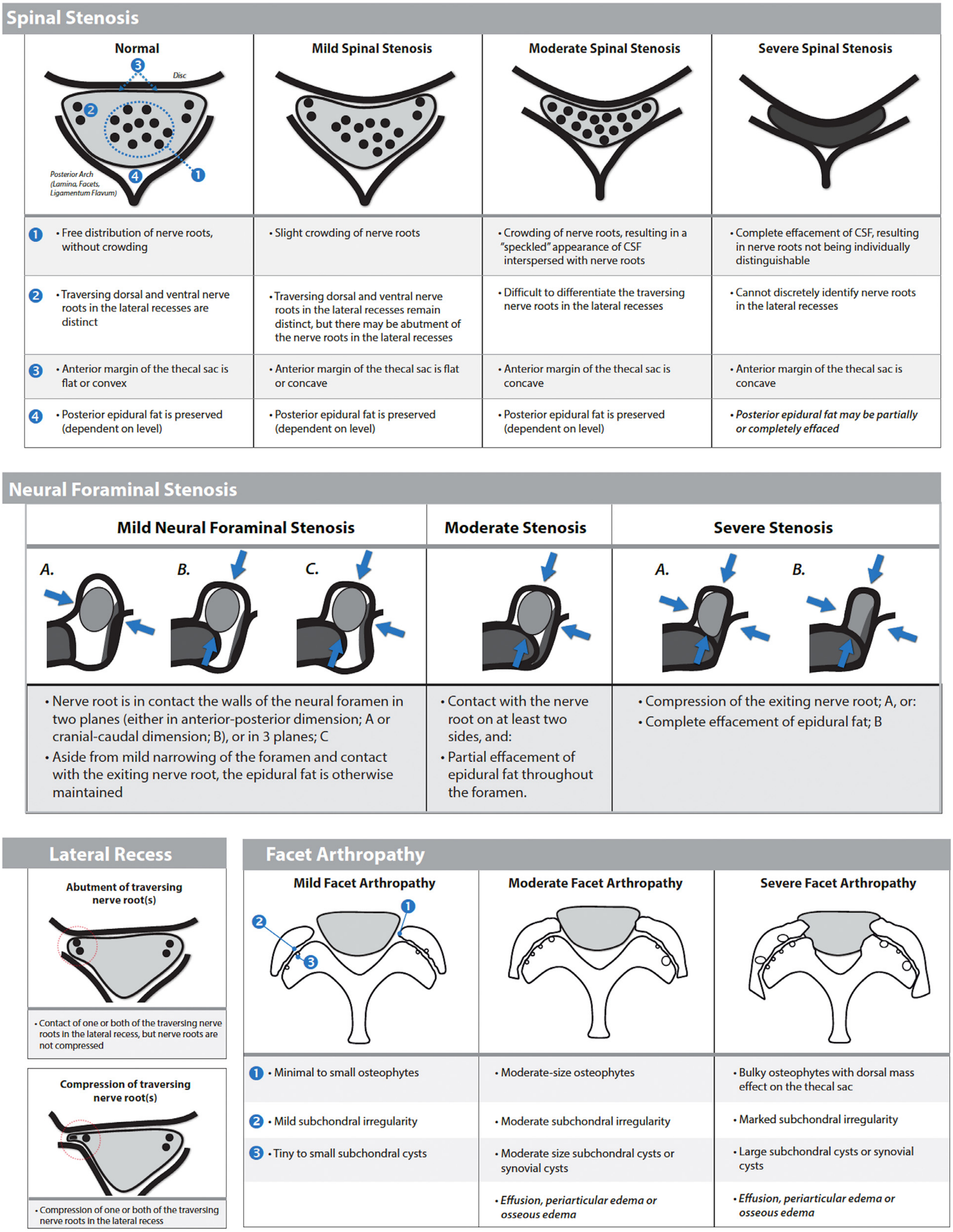

FIG 1. Schematic summary of the MRI-based standardized classification for lumbar spine degeneration on MRI. Miskin N, Isaac Z, Lu Y et al. *Pain Medicine*, 2021, Volume 22, Issue 7, Pages 1485-1495, by permission of Oxford University Press.

classifications. For the assessment of NFS, the MSK group demonstrated substantial agreement both before training (k=0.668; 95% confidence interval (CI) 0.623-0.713) and after training (k=0.621; CI 0.577-0.666). The NR group demonstrated moderate agreement both before training (k=0.541; CI 0.489-0.593) and after training (k=0.576; CI 0.527-0.626). The IS analysis demonstrated an improvement in moderate agreement before training (k=0.527; CI 0.498-0.556) to substantial agreement after training (k=0.602; CI

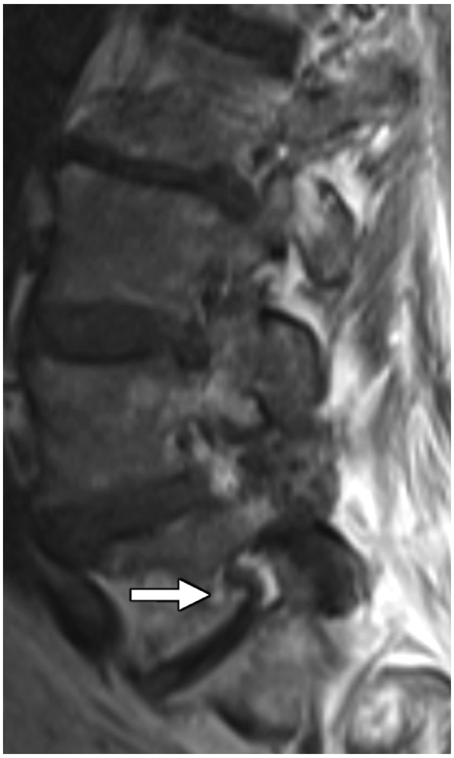

FIG 2. Sagittal T1-weighted image at the level of the exiting left L5 nerve root (arrow). Prior to training, the six readers described the degree of neural foraminal stenosis, with one reporting it as mild, three reporting it as mild-moderate, one reporting it as moderate, and one reporting it as moderate-severe. After training, five reported it as moderate and one reported it as severe.

0.576-0.629). Of note, these 95% confidence intervals do not overlap. An example of improved concordance in the reporting of NFS is shown in Figure 2.

Spinal Canal Stenosis

One hundred assessments for SCS were obtained (two levels for the 50 exams). For the assessment of SCS, the MSK group demonstrated substantial agreement before training (k=0.645; CI 0.554-0.736) and moderate agreement after training (k=0.536; CI 0.424-0.649). The NR group demonstrated an improvement in fair agreement before training (k=0.368; CI 0.261-0.474) to substantial agreement after training (k=0.638; CI 0.568-0.709) with nonoverlapping 95% confidence intervals. The IS analysis demonstrated an improvement in moderate agreement before training (k=0.540; CI 0.486-0.594) to substantial agreement after training (k=0.652; CI 0.612-0.693), with nonoverlapping 95% confidence intervals. An example demonstrating improved concordance for the reporting of SCS in the NR group and persistent variability in reporting in the MSK group is shown in Figure 3.

Lateral Recess Stenosis

Two hundred lateral recesses (two sides at two levels for the 50 exams) were assessed. For the assessment of LRS, the MSK group demonstrated an improvement in poor agreement before training (k= -0.0101; CI -0.577 to 0.557) to slight agreement after training (k=0.136; CI 0.0552-0.216). The NR group demonstrated an improvement in fair agreement before training (k=0.392; CI 0.267-0.516) to moderate agreement after training (k=0.480; CI 0.418-0.541). The IS analysis demonstrated an improvement in slight agreement before training (k=0.0818; CI -0.0323 to 0.196) to fair agreement after training (k=0.337; CI 0.296-0.378), with nonoverlapping 95% confidence intervals. An example showing improved concordance in the reporting of LRS is shown in Figure 4.

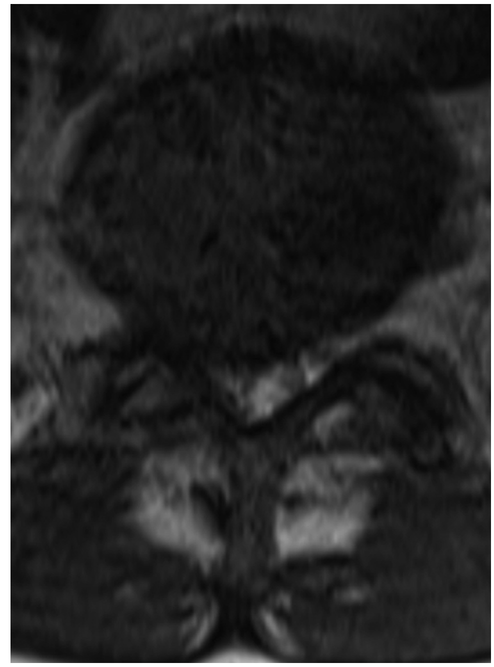

FIG 3. Axial T2-weighted image at the level of the L5-S1 disc. Prior to training, the three neuroradiologist readers described the spinal canal stenosis (SCS) differently, as none, moderate, and moderate-severe. After training, their assessments were moderate, moderate, and moderate-severe. Prior to training, the three musculoskeletal radiologists described the SCS as mild, moderate, and moderate-severe, and after training assessments remained variable: mild-moderate, moderate-severe, and severe.

FIG 4. Axial T2-weighted image at the level of the L4-L5 disc. Prior to training, all three neuroradiologist readers reported right lateral recess narrowing (arrows), and none of the musculoskeletal radiologists reported this finding. After training, four of the six readers (two each from the NR and MSK groups) described lateral recess narrowing, describing it as nerve root abutment.

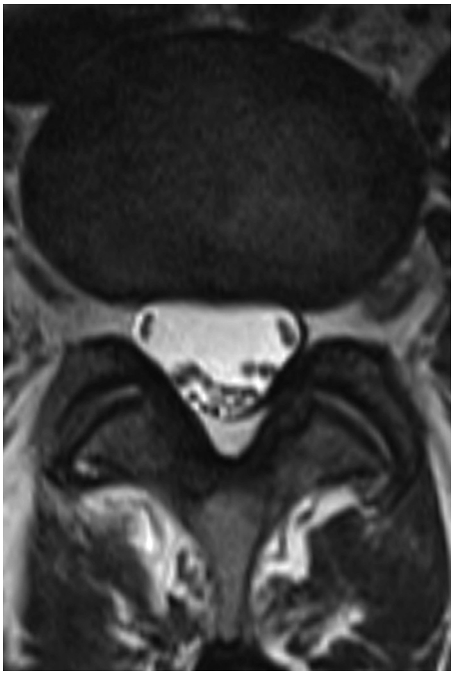

FIG 5. Axial T2-weighted image at the level of the L4-L5 disc. Prior to training, only two MSK readers of the total six readers reported the presence of facet osteoarthritis. After training, all readers but one of the NR readers reported facet osteoarthritis, all of which described it as mild in severity.

Facet Osteoarthritis

One hundred pairs of facet joints (two levels for the 50 fifty exams) were assessed. For the assessment of FO, the MSK group demonstrated an improvement in slight agreement before training (k=0.134; CI -0.0548 to 0.323) to moderate agreement after training (k=0.413; CI 0.345-0.481), with nonoverlapping 95% confidence intervals. The NR group demonstrated an improvement in fair agreement before training (k=0.359; CI 0.248-0.470) to moderate agreement after training (k=0.419; CI 0.348-0.490). The IS analysis demonstrated an improvement in slight agreement before training (k=0.176; CI 0.100-0.252) to moderate agreement after training (k=0.495; CI 0.455-0.534), with nonoverlapping confidence intervals. An example demonstrating the improved concordance in the reporting of FO is shown in Figure 5.

Summary of intra-subspecialty and inter-subspecialty kappa for the parameters prior to after training is shown in Figure 6.

Discussion

The results demonstrate improved inter-subspecialty agreement for all four parameters after training with the standardized classification system of degenerative change, with nonoverlapping confidence intervals (from moderate to substantial for NFS and SCS, from slight to fair for LRS, from slight to moderate for FO). There is improved intra-subspecialty agreement after training in the NR group for the assessment of SCS (from fair to substantial), and the MSK group for the assessment of FO (from slight to moderate), with nonoverlapping 95% confidence intervals. There is slightly decreased agreement in the MSK group for the assessment of NFS and SCS, however the 95% confidence intervals overlap and are thus not significant. We demonstrate similar intra-subspecialty agreement in both groups for all other parameters.

Prior studies have also sought to reduce variability in lumbar MRI reporting with classification systems, which have been tested in a group of radiologists. However, it is important to note that to our knowledge, no classification system has sought to specifically reduce radiologist inter-subspecialty variability as we have shown in this study. Lee et al[17] created a classification system, later re-tested by

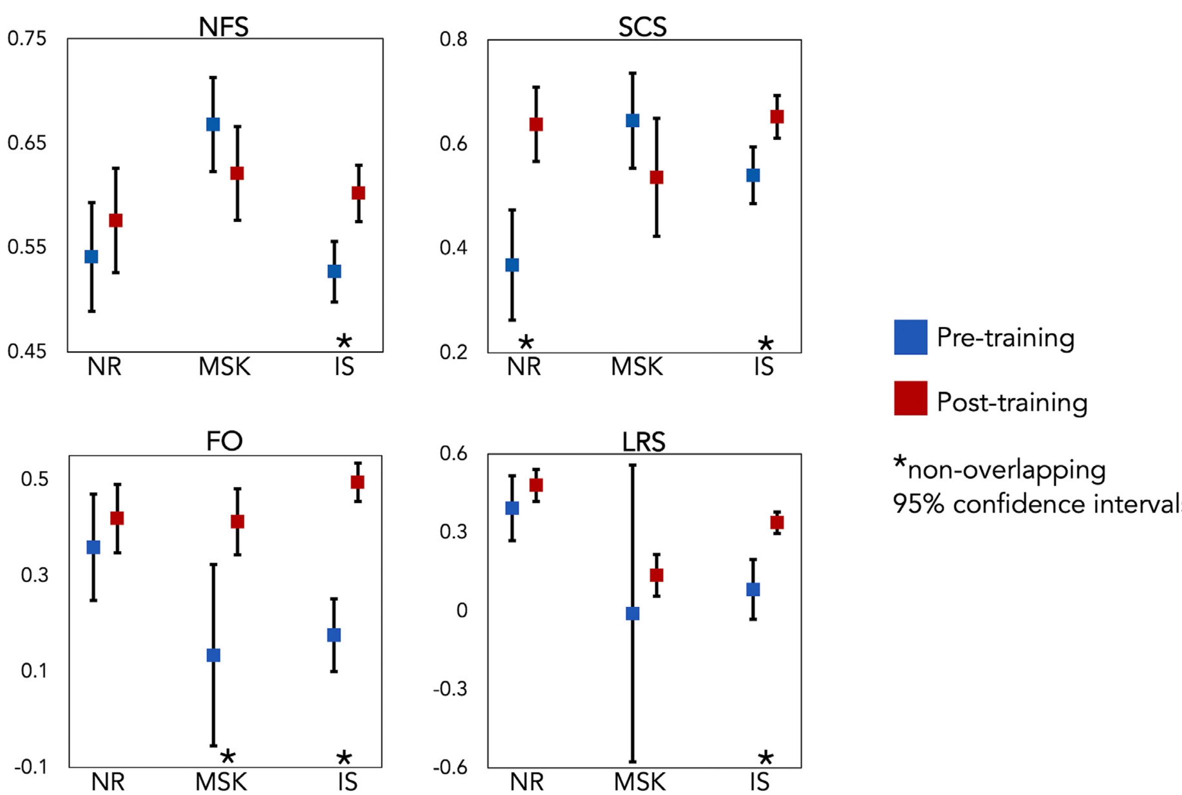

FIG 6. Representation of Cohen's kappa coefficient with 95% confidence intervals (y-axis) for each of the parameters assessed for concordance, both prior to and after training. FO, facet osteoarthritis; IS, inter-subspecialty; LRS, lateral recess stenosis; MSK, musculoskeletal radiologist group; NFS, neural foraminal stenosis; NR, neuroradiologist group; SCS, spinal canal stenosis.

Park et al[18] to reduce variability in the reporting of neural foraminal stenosis, and similarly Berg et al[19] developed a classification to assess variability in reporting facet osteoarthritis. These studies focused their efforts to reduce variability in radiology reporting by evaluating a single parameter of degeneration. However, our system assesses the reporting of four components comprising the clinically important parameters assessed on lumbar spine MRI. Furthermore, the current study differs in that it allows for comparison of reporting prior and after training with a system.

More recently, Wang et al[9] demonstrated increased agreement in reports generated with the aid of a computer-assisted reporting tool, as compared to free-text reports in a group of five MSK radiologists. Direct comparison with the current study is difficult due to different statistical methods. Furthermore, this study did not seek to define what "mild" spinal canal stenosis entailed, but rather, this tool confined reader interpretation of degree of stenosis to be reported as one of four ordinal categories. The efficacy of this tool was also not validated with a group of neuroradiologists, and no comparison was made between MSK and NR groups.

Fu et al[5] also developed a standardized classification of degenerative change, demonstrating improvement in variability, a group of MSK radiologists and orthopedic surgeons. Again, direct comparison with the current study is limited due to different statistical methods. This study however did include standardized assessments of multiple parameters assessed in lumbar degeneration, but limited definitions to descriptions alone, not schematic depictions or patient examples as in this study. This study also differs from ours in that the efficacy of the classification is unknown since baseline comparison of free-text reports prior to providing these standardized assessments was not performed. Moreover, the previously described study, this assessment guide was not validated with a group of neuroradiologists.

Our study has several limitations. It is possible that after informing the readers of the parameters assessed in the first part of our study that this alone would modify their reporting of degenerative change alone to fixate on these four parameters, which is perhaps a reason to explain any improvements after re-reading the exams. However, the results are based on concordance in grading the *severity* of degenerative change, not merely reporting that a given feature of degeneration is present or absent, suggesting that training with the standardized classification is the reason for the findings. A related limitation is the comparison of pre-training kappa values for LRS and FO, in which concordance consisted of denoting the presence of absence of reporting, compared with post-training kappa values for LRS and FO, which were based on an ordinal scale. We suspect that the use of this ordinal scale post-training potentially underestimates the relative improvements demonstrated in intra- and inter-subspecialty variability.

An important limitation is the absence of a gold standard in the reporting of lumbar spine degeneration. Despite the fact that these standardized definitions were developed with multidisciplinary consensus, they are a schematic or guide to aid in classification and for a given vertebral body interspace for a given patient there is no ground truth in the quantification of degeneration. This study focused on increased concordance as the indicator of success of the classification. The extent to which a reduction in variability translates to any improvement in patient outcomes is unknown, and is a potential topic for future work, potentially with the aid of this classification to investigate if the severity distinctions made in this classification system translate to improvement after conservative or procedural interventions.

Conclusion

A standardized classification of degenerative change in the lumbar spine decreased intra-subspecialty variability in the grading of SCS in the NR group, the grading of FO in the MSK group, and inter-subspecialty variability of all four parameters (NFS, SCS, LRS, FO) between NR and MSK groups. The results suggest this classification may be a clinically useful tool to improve the consistency of reporting lumbar spine MRI exams for both subspecialties.

References

1. Patel ND, Broderick DF, Burns J, et al. ACR appropriateness criteria low back pain. J Am Coll Radiol JACR 2016;13:1069–78.
2. Chou R, Qaseem A, Snow V, et al. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med 2007;147:478–91.
3. Akuthota V, Meron AJ, Singh JR, et al. The utility of magnetic resonance imaging results in physician decision-making before initial lumbar spinal injection. Spine J Off J North Am Spine Soc 2019;19:1455–62.
4. El Barzouhi A, Vleggeert-Lankamp CLAM, Lycklama À Nijeholt GJ, et al. Magnetic resonance imaging interpretation in patients with sciatica who are potential candidates for lumbar disc surgery. PloS One 2013;8:e68411.
5. Fu MC, Buerba RA, Long WD, et al. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. Spine J Off J North Am Spine Soc 2014;14:2442–8.
6. Winklhofer S, Held U, Burgstaller JM, et al. Degenerative lumbar spinal canal stenosis: intra- and inter-reader agreement for magnetic resonance imaging parameters. Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc 2017;26:353–61.
7. Mulconrey DS, Knight RQ, Bramble JD, et al. Interobserver reliability in the interpretation of diagnostic lumbar MRI and nuclear imaging. Spine J Off J North Am Spine Soc 2006;6:177–84.
8. van Rijn JC, Klemetsö N, Reitsma JB, et al. Observer variation in MRI evaluation of patients suspected of lumbar disk herniation. AJR Am J Roentgenol 2005;184:299–303.
9. Wang B, Rosenthal DI, Xu C, et al. The effect of computer-assisted reporting on interreader variability of lumbar spine MRI degenerative findings: five readers with 30 disc levels. J Am Coll Radiol JACR 2018, https://doi.org/10.1016/j.jacr.2017.12.020.
10. Cihangiroglu M, Yildirim H, Bozgeyik Z, et al. Observer variability based on the strength of MR scanners in the assessment of lumbar degenerative disc disease. Eur J Radiol 2004;51:202–8.
11. Speciale AC, Pietrobon R, Urban CW, et al. Observer variability in assessing lumbar spinal stenosis severity on magnetic resonance imaging and its relation to cross-sectional spinal canal area. Spine 2002;27:1082–6.
12. Moojen WA, Schenck CD, Lycklama À Nijeholt GJ, et al. Preoperative MRI in patients with intermittent neurogenic claudication: relevance for diagnosis and prognosis. Spine 2018;43:348–55.
13. 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;49:182–7.
14. Miskin N, Isaac Z, Lu Y, et al. Simplified Universal Grading of Lumbar Spine MRI Degenerative Findings: Inter-reader agreement of non-radiologist spine experts. Pain Med Malden Mass 2021, https://doi.org/10.1093/pm/pnab098.
15. Hostetter J, Khanna N, Mandell JC. Integration of a Zero-footprint Cloud-based picture archiving and communication system with customizable forms for radiology research and education. Acad Radiol 2018;25:811–8.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
17. Lee S, Lee JW, Yeom JS, et al. A practical MRI grading system for lumbar foraminal stenosis. AJR Am J Roentgenol 2010;194:1095–8.
18. Park H-J, Kim SS, Lee S-Y, et al. Clinical correlation of a new MR imaging method for assessing lumbar foraminal stenosis. AJNR Am J Neuroradiol 2012;33:818–22.
19. Berg L, Thoresen H, Neckelmann G, et al. Facet arthropathy evaluation: CT or MRI? Eur Radiol 2019;29:4990–8.