ORIGINAL RESEARCH

Influence of Nomenclature in the Interpretation of Lumbar Disk Contour on MR Imaging: A Comparison of the Agreement Using the Combined Task Force and the Nordic Nomenclatures

E. Arana
F.M. Kovacs
A. Royuela
A. Estremera
H. Sarasíbar
G. Amengual
I. Galarraga
C. Martínez
A. Muriel
V. Abraira
J. Zamora
C. Campillo

FELLOWS' JOURNAL CLUB

BACKGROUND AND PURPOSE: The CTF nomenclature had not been tested in clinical practice. The purpose of this study was to compare the reliability and diagnostic confidence in the interpretation of disk contours on lumbar 1.5T MR imaging when using the CTF and the Nordic nomenclatures.

MATERIALS AND METHODS: Five general radiologists from 3 hospitals blindly and independently assessed intravertebral herniations (Schmorl node) and disk contours on the lumbar MR imaging of 53 patients with low back pain, on 4 occasions. Measures were taken to minimize the risk of recall bias. The Nordic nomenclature was used for the first 2 assessments, and the CTF nomenclature, in the remaining 2. Radiologists had not previously used either of the 2 nomenclatures. κ statistics were calculated separately for reports deriving from each nomenclature and were categorized as almost perfect (0.81–1.00), substantial (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), slight (0.00–0.20), and poor (<0.00).

RESULTS: Categorization of intra- and interobserver agreement was the same across nomenclatures. Intraobserver reliability was substantial for intravertebral herniations and disk contour abnormalities. Interobserver reliability was moderate for intravertebral herniations and fair to moderate for disk contour.

CONCLUSIONS: In conditions close to clinical practice, regardless of the specific nomenclature used, a standardized nomenclature supports only moderate interobserver agreement. The Nordic nomenclature increases self-confidence in an individual observer's report but is less clear regarding the classification of disks as normal versus bulged.

ABBREVIATIONS: CTF = Combined Task Force of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology; Nordic = Nordic Modic Consensus Group Classification; Rx = radiologist

SPINE

ORIGINAL RESEARCH

The interpretation of lumbar spine MR images often has a great influence on diagnosis, treatment, and prognosis of low back pain.[1] However, evidence-based guidelines recommend lumbar MR imaging only when "red flags" on clinical history or physical examination suggest that pain may be caused by certain systemic diseases (such as cancer) or when clinical indication criteria for surgery exist (such as irradiated pain lasting for >6–12 weeks attributed to disk herniation) because correlation between clinical features and MR imaging findings helps to confirm the diagnosis and management.[2-4]

One of the reasons for limiting MR imaging to such cases is that the interobserver agreement in the interpretation of lumbar MR imaging is, at best, only moderate.[5-10] Some studies may have overestimated the agreement because they analyzed only reports from 2 or 3 expert readers who were specialists in the area of lumbar MR imaging and worked in a single research setting,[6,7,11,12] which may have led to a formal or informal agreement in their diagnostic criteria. Agreement, to be expected in routine clinical practice, may be better estimated by analyzing reports from a higher number of radiologists with no prior formal or informal consensus on diagnostic standards.[8,9]

It has been suggested that the use of ambiguous nomenclature may be a reason for this agreement being relatively low.[2,12,13] Two of the most commonly used nomenclature systems for degenerative disk disease are the CTF, endorsed by the American Society of Neuroradiology, among other societies,[14] and the Nordic.[15,16] The latter is based on previously

Received August 29, 2010; accepted after revision November 1.

From the Department of Radiology, (E.A.), Fundación Instituto Valenciano de Oncología, Valencia, Spain; Fundación Instituto de Investigación en Servicios de Salud (E.A.) Valencia, Spain; Spanish Back Pain Research Network (E.A., F.M.K., A.R., A.E., H.S., G.A., I.G., C.M., A.M., V.A., J.Z., C.C.) and Departamento Científico (F.M.K.), Fundación Kovacs, Palma de Mallorca, Spain; Centro de investigación Biomédica en Red, Epidemiología y Salud Pública (A.R., A.M., V.A., J.Z.), Barcelona, Spain; Unidad de Bioestadística Clínica (A.R., A.M., V.A., J.Z.), Hospital Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain; Hospital Son Llàtzer (A.E., H.S., G.A., C.M.), Palma de Mallorca, Spain; Hospital de Manacor (I.G.), Mallorca, Spain; and Ib-Salut (C.C.), Palma de Mallorca, Spain

This work was supported by the Kovacs Foundation, Palma de Mallorca, Spain. The Kovacs Foundation is a Spanish not-for-profit research institution, with no commercial activity or links to the health industry. Its Board of Trustees includes the Spanish Minister of Health, Spanish Red Cross, the Spanish Medical Association, and other governmental and private institutions (www.kovacs.org). The following individuals received research support, including provision of equipment or materials: Estanislao Arana, Francisco Kovacs, Ana Estremera, Helena Sarasíbar, Isabel Galarraga, Alfonso Muriel, Javier Zamora, and Carlos Campillo Artero.

Please address correspondence to Estanislao Arana, MD, MHA, PhD, Department of Radiology, Fundación Instituto Valenciano de Oncología, C/ Beltrán Báguena, 19, 46009 Valencia, Spain; e-mail: aranae@uv.es

indicates article with supplemental on-line tables.

DOI 10.3174/ajnr.A2448

existing terminology[7] and leads to only moderate interobserver agreement in the interpretation of lumbar MR imaging.[10] Although the agreement between experts participating in a clinical trial has been calculated for some of the items of the CTF,[12] the agreement derived from its use among community radiologists in conditions similar to routine clinical practice is currently unknown. Comparison of such agreement with that deriving from the use of the Nordic nomenclature has recently been recommended.[17]

Our objective was to prospectively compare the reliability and diagnostic confidence in the interpretation of disk contour on lumbar 1.5T MR images deriving from the use of the CTF and Nordic nomenclatures.

Materials and Methods

This prospective study was approved by the institutional review boards of the participating hospitals.

Study Population

Five practicing general radiologists (E.A., A.E., H.S., G.A., I.G.), working in 3 general hospitals located in 2 different geographic regions, participated in this study. Their postresidency experience as radiologists ranged from 12 to 18 years, and their experience interpreting spine imaging ranged from 8 to 12 years. They were trained in different institutions without fellowships.

Two of the radiologists, working in hospitals in different cities, randomly selected images from 68 patients who had undergone 1.5T MR imaging for low back pain and/or sciatica. Exclusion criteria were the following: previous spine surgery, pregnancy, cauda equina syndrome, scoliosis with >15° curvature, vertebral fractures, inflammatory spondyloarthropathy, spinal infection, or tumor. Exclusions were the following: 7 patients for previous spine surgery, 5 for scoliosis, and 3 for metastatic cancer, leaving a total sample of 28 women and 25 men. The mean age for men and women was similar (46.3 ± 13.7 years and 50.3 ± 12.9 years, respectively; $P = .274$). Images in this study were those used to assess agreement derived from the Nordic nomenclature.[9]

MR Imaging

All examinations were performed on two 1.5T MR imaging systems with a 6-channel phased-array spine coil. The diagnostic imaging protocol consisted of an MR study of the lumbar spine, standardized as shown in On-Line Table 1 without fat suppression. All images were unlabeled, for patient confidentiality and radiologist masking, as to data on sex and age, and were distributed to all radiologists participating in this study.

Variables

Radiologists reported their findings by using the Nordic and the CTF forms.[14,16] All the variables were recorded separately for all the lumbar segments (from L1-L2 to L5-S1).

Among the variables assessed on the Nordic form, those analyzed for this study were Schmorl nodes (yes/no) and disk contour (normal, bulging, protrusion [focal or broad-based], and hernia [extrusion or sequestration]). With the CTF form, the recorded variables were intravertebral herniation (yes/no) and disk contour (normal, symmetric bulging disk, focal-based herniation, broad-based herniation, and extrusion).

In addition, the degree of confidence of the reader with respect to the classification of each image was also gathered by using a 3-point scale.[14] Possible values for each diagnosis were "definite" (no doubt), "probable" (some doubt, but likelihood >50%), or "possible" (some reason to consider the diagnosis, but likelihood <50%). At the analysis stage, this scale was collapsed into 2 categories: "likely" (definite + probable) versus "possible."

Assessments and Data Collection

All MR images were presented on compact discs created by using K-PACS imaging software, Version V0.9.5.3; (IMAGE Information Systems, Plauen, Germany). The types and numbers of display monitors used were not standardized among the readers.

The 5 radiologists were unaware of any demographic and clinical features of the patients from whom the images had been taken. They were asked to report their findings independent of their opinion on the clinical relevance of those findings with closed-ended responses, by using the Nordic and CTF nomenclatures. Participating radiologists were untrained in both nomenclatures and were only provided with the definitions included in the CTF and Nordic forms, as shown in On-Line Table 2. Otherwise, they were asked to act as they usually do in their routine clinical practice. No attempt was made to further define or standardize the meaning of each term or to homogenize the diagnostic criteria, and they received no instructions regarding the interpretation of images. They assessed the MR images alone and on their own.

They first assessed the MR images by using the Nordic form. To assess intraobserver reliability, the 5 radiologists were asked to reevaluate the same MR images with the same form, at least 14 days after the forms with their first interpretation had been collected.

Radiologists were unaware that the images they assessed at the second round were the same.

At least 6 months later, the same procedure was followed by using the CTF form. The radiologists were asked to assess the same set of MR images twice, with a minimum of 14 days' interval. Therefore, the same set of images was interpreted by each radiologist 4 times, twice with Nordic and twice with CTF. The images were presented in a different order each time, and radiologists had no access to their previous reports or to their colleagues' current or previous reports.

All reports were entered in the data base at a centralized coordination office. Entry of data was done independently by 2 administrative assistants, who double-checked that the data they were entering coincided with the information on the forms.

Data Analysis

Ratings from each observer were cross-tabulated, and intra- and interobserver agreements were measured by using the κ statistic. Two sets of analyses were done, each deriving from reports using the Nordic and the CTF nomenclatures, respectively.

κ values were categorized as reflecting an almost perfect (0.81–1.00), substantial (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), slight (0.00–0.20), or poor (<0.00) agreement.[18]

The κ statistic is influenced by the prevalence of the events, so that findings with very high or very low prevalence lead to very low κ values, even if the observer agreement is high.[18,19] Therefore, at the design phase, it was decided that κ values would be calculated only for findings reported in >10% and in <90% of reports. Five radiologists interpreted 53 images (total, 265 reports). Hence, κ values were not calculated for findings identified in ≤27 or in ≥238 of those reports.

The unit of analysis was imaging at each disk level. To make it possible to calculate the κ statistic, we dichotomized reports into only 2 categories ("normal" or "abnormal"). It was necessary to collapse

Table 1: Radiologic findings per vertebral level

Variable	L1-L2	L2-L3	L3-L4	L4-L5	L5-S1
CTF nomenclature					
Intravertebral herniation (yes)[a]	64 (24.2)	60 (22.6)	58 (21.9)	53 (20.0)	56 (21.1)
Disk contour[a]					
Normal	225 (84.9)	183 (69.1)	137 (51.7)	64 (24.2)	112 (42.3)
Bulging (>50%)	30 (11.3)	67 (25.3)	108 (40.8)	136 (51.3)	72 (27.2)
Protrusion focal (<25%)	8 (3.0)	6 (2.3)	16 (6.0)	53 (20.0)	47 (17.7)
Protrusion broad-based (25%–50%)	2 (0.8)	5 (1.9)	4 (1.5)	5 (1.9)	22 (8.3)
Extrusion	0 (0.0)	4 (1.5)	0 (0.0)	7 (2.6)	12 (4.5)
Nordic nomenclature					
Intravertebral herniation (yes)[a]	50 (18.9)	46 (17.4)	52 (19.6)	36 (13.6)	24 (9.1)
Disk contour[a]					
Normal	217 (81.9)	204 (77.0)	160 (60.4)	84 (31.7)	104 (39.2)
Bulging	45 (17.0)	46 (17.4)	97 (36.6)	137 (51.7)	98 (37.0)
Protrusion	3 (1.1)	11 (4.2)	8 (3.0)	39 (14.7)	60 (22.6)
Extrusion	0 (0.0)	4 (1.5)	0 (0.0)	5 (1.9)	3 (1.1)

[a] No. (%). All of the percentages are calculated over a total of 265 images (53 images seen by the 5 radiologists).

Table 2: Intraobserver agreement

Variable	Intraobserver Agreement CTF[a]	Intraobserver Agreement Nordic[a]
Intravertebral herniation	0.682 (0.389; 0.835)	0.694 (0.530; 0.852)
Disk contour at L1-L2, L2-L3, and L3-L4[b]	0.750 (0.654; 0.824)	0.714 (0.608; 0.776)
Disk contour at L4-L5 and L5-S1		
Considering bulging as abnormal [b]	0.728 (0.622; 0.864)	0.642 (0.558; 0.816)
Considering bulging as normal[c]	0.624 (0.538; 0.667)	0.565 (0.460; 0.717)

[a] Mean (5th percentile, 95th percentile) of κ values. Agreement is classified as almost perfect (κ value $>$ 0.81), substantial (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), slight (0.00–0.20), or poor (<0.00).
[b] Agreement in classifying images in the 2 following categories: normal vs bulging + focal protrusion + broad-based protrusion + extrusion, according to the CTF nomenclature; or normal vs symmetric bulging + protrusion (focal or broad-based) + hernia (extrusion or sequestration), according to the Nordic nomenclature.
[c] Agreement in classifying images in the 2 following categories: normal + bulging vs focal protrusion + broad-based protrusion + extrusion, according to the CTF nomenclature; or normal + symmetric bulging vs protrusion (focal or broad-based) + hernia (extrusion or sequestration), according to the Nordic nomenclature.

some categories due to a prevalence of findings below 10%. As a result, it would have been inappropriate to use a weighted-κ approach, and a mean κ pair-wise comparison was undertaken. For disk contour, "normal" was considered to be a normal disk, and "abnormal," all the other findings (bulging, focal protrusion, broad-based protrusion, and extrusion, according to the CTF nomenclature, or bulging, protrusion, and extrusion, according to the Nordic nomenclature).

Findings at each level were listed, and those for which there was a prevalence between 10% and 90% were identified. Because there was a correlation between the findings at different vertebral levels on the same image, κ was calculated following the 2-step approximation described by Lipsitz et al.[20] This approximation essentially consists of estimating the expected and observed probabilities by means of logistic regression. In this case, the regression model included vertebral level, age, sex, and the interaction between age and sex. Generalized estimating equation models were used.[21] The structure was a self-regressive correlation.

Statistical packages (STATA IC/10.0 for Windows, StataCorp, College Station, Texas; and Statistical Package for the Social Sciences, Version 16.0, SPSS, Chicago, Illinois) were used for data analysis.

Results

Most findings related to disk contour were found at the L4-L5 and L5-S1 levels, while most intravertebral herniations (Schmorl nodes) were reported at the L1-L2 and L2-L3 levels (Table 1). The prevalence of findings varied depending on the nomenclature used (Table 1). The prevalence of findings did not permit statistically sound comparisons in reproducibility at individualized levels, except at L4-L5 and L5-S1 levels. At these disks, the number of reports on abnormal findings on disk contour allowed exploring a different definition for normal, which included bulging as well.

With either of the 2 nomenclatures, intraobserver agreement was substantial for all findings. However, intraobserver agreement by using the Nordic nomenclature was only moderate when bulging was categorized as normal. The intraobserver agreement derived from the use of both nomenclatures was very similar for all findings, though κ values were slightly lower for the Nordic nomenclature (Table 2).

The mean intraobserver agreement of all reports (with both nomenclatures and all definitions for normality) was 0.674 (5th percentile, 95th percentile: 0.634, 0.738). Independent of the nomenclature used, the overall intraobserver agreement was very similar across radiologists (data not shown). The mean (5th percentile, 95th percentile: κ value for the radiologist with the lowest intraobserver agreement was 0.603 (0.438, 0.743); that for the radiologist with the highest agreement was 0.729 (0.672, 0.816). Both values fell in the "substantial" category.

Interobserver agreement was moderate in all findings for both nomenclatures, except for disk contour, in which bulging was categorized as normal and agreement was only fair. In this case, the range of agreement was wider by using the Nordic than the CTF nomenclature when categorization depended on whether bulging was considered normal (Table 3). The interobserver agreement derived from the use of both nomenclatures was very similar, though κ values were consistently

Table 3: Number of images classified by each radiologist as normal or abnormal using both nomenclatures and depending on whether bulging was considered to be normal or abnormal (levels L4-L5 and L5-S1)

	CTF								Nordic							
	Control 1				Control 2				Control 1				Control 2			
	Considering Bulging Abnormal[a]		Considering Bulging Normal[b]		Considering Bulging Abnormal[a]		Considering Bulging Normal[b]		Considering Bulging Abnormal[a]		Considering Bulging Normal[b]		Considering Bulging Abnormal[a]		Considering Bulging Normal[b]	
	N	A	N	A	N	A	N	A	N	A	N	A	N	A	N	A
L4-L5																
Rx1	5	48	32	21	2	51	30	23	11	42	32	21	13	40	32	21
Rx2	5	48	48	5	3	50	43	10	9	44	49	4	9	44	49	4
Rx3	13	40	37	16	14	39	43	10	21	32	42	11	25	28	39	14
Rx4	23	30	47	6	19	34	44	9	21	32	48	5	29	24	46	7
Rx5	18	35	36	17	22	31	36	17	22	31	50	3	23	30	52	1
L5-S1																
Rx1	24	29	34	19	20	33	33	20	21	32	29	24	23	30	34	19
Rx2	16	37	42	11	11	42	46	7	11	42	49	4	18	35	50	3
Rx3	20	33	30	23	22	31	39	14	20	33	37	16	25	28	33	20
Rx4	28	25	44	9	24	29	40	13	28	25	43	10	35	18	44	9
Rx5	24	29	34	19	25	28	33	20	24	29	44	9	26	27	48	5

[a] Agreement in classifying images in the 2 following categories: normal vs bulging + focal protrusion + broad-based protrusion + extrusion, according to the CTF nomenclature; or normal vs symmetric bulging + protrusion (focal or broad-based) + hernia (extrusion or sequestration), according to the Nordic nomenclature.
[b] Agreement in classifying images in the 2 following categories: normal + bulging vs focal protrusion + broad-based protrusion + extrusion, according to the CTF nomenclature; or normal + symmetric bulging vs protrusion (focal or broad-based) + hernia (extrusion or sequestration), according to the Nordic nomenclature.

Table 4: Interobserver agreement

Variable	Interobserver Agreement CTF[a]	Interobserver Agreement Nordic[a]
Intravertebral herniation	0.530 (0.415; 0.657)	0.481 (0.236; 0.673)
Disk contour at L1-L2, L2-L3, and L3-L4[b]	0.476 (0.398; 0.573)	0.473 (0.286; 0.560)
Disk contour at L4-L5 and L5-S1		
Considering bulging as abnormal[b]	0.562 (0.426; 0.788)	0.456 (0.179; 0.650)
Considering bulging as normal[c]	0.407 (0.261; 0.597)	0.277 (0.000; 0.571)

[a] Mean (5th percentile, 95th percentile) of κ values. Agreement is classified as almost perfect (κ value $>$ 0.81), substantial (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), slight (0.00–0.20), or poor ($<$0.00).[18]
[b] Agreement in classifying images in the 2 following categories: normal vs bulging + focal protrusion + broad-based protrusion + extrusion, according to the CTF nomenclature; or normal vs symmetric bulging + protrusion (focal or broad-based) + hernia (extrusion or sequestration), according to the Nordic nomenclature.
[c] Agreement in classifying images in the 2 following categories: normal + bulging vs focal protrusion + broad-based protrusion + extrusion, according to the CTF nomenclature; or normal + symmetric bulging vs protrusion (focal or broad-based) + hernia (extrusion or sequestration), according to the Nordic nomenclature.

slightly lower for the Nordic nomenclature (Table 4). As for the diagnosis of extrusion, only 5% of κ scores were above 0.657 (by using the CTF nomenclature) or 0.673 (by using the Nordic nomenclature).

The degree of radiologists' confidence in their own reports was higher when they used the Nordic than when they used the CTF nomenclature. When using the Nordic nomenclature the first time they assessed the images, they rated 96.2% of their findings on disk contour and 98.9% of those on Schmorl nodes as "likely." In the second round, these proportions rose to 97.4% and 99.6%, respectively. Corresponding proportions when using the CTF nomenclature were 73.2% and 91.7% in the first round, but they decreased to 63.4% and 74.3% in the second round.

Discussion

A previous study had assessed the agreement derived from the use of this nomenclature but only among a few experts interpreting MR images from a sample of patients selected among those included in a clinical trial.[14] To the authors' knowledge, the current study is the first to assess the reliability of the CTF nomenclature in a community setting, by using the conditions that are recommended for studies on agreement.[22]

In the current study, κ values reflecting both intra- and interobserver agreement for intravertebral herniation and disk contour were slightly higher with the CTF than with the Nordic nomenclature. However, the category of interobserver agreement was the same regardless of the nomenclature used. This finding questions the clinical relevance of the differences found in the agreement between these nomenclatures.

Conversely, the radiologists' confidence in their own assessments was higher when they used the Nordic nomenclature. Moreover, the degree of confidence in their own diagnoses increased between the first and second assessment when they followed the Nordic nomenclature, while it decreased when they followed the CTF nomenclature. This might be related to participating radiologists' lack of confidence in assessing whether a protrusion reached 25% of the disk circumference which, according to the CTF nomenclature, determines whether it is focal or broad-based. The fact that the CTF nomenclature led to a higher prevalence of findings may also account for the decrease in radiologists' confidence.[23] In any case, it should be noted that diagnostic confidence is not a reliable measure of diagnostic accuracy.[24]

A bulging disk at L4-L5 or, especially at L5-S1, can be the consequence of disk degeneration but is usually a clinically irrelevant normal variation.[2,11,25] In this study, both nomenclatures yielded better interobserver agreement when bulging

was included in the "abnormal" category. In fact, the "bulging" category is the main reason for disagreement,[2,7] probably because radiologists use it as an escape option when in doubt.[26]

Radiologists who participated in this study worked in different hospitals, did not have a fellowship, had not met previously to agree on diagnostic criteria, had not received any specific training with example case sets, did not use templates or on-line examples, and were unaware of patients' clinical features; and no effort was implemented to standardize nomenclature, apart from definitions included in the Nordic and CTF forms, as recently recommended.[22] In theory, all these features might have lowered interobserver agreement.[6,16,22,27] However, previous studies conducted in different settings consistently reported agreement in the interpretation of lumbar MR imaging to be only moderate, even though radiologists were highly trained experts, templates were used, or radiologists tried to reach consensus in previous meetings.[6-8,12,26,28] Moreover, the goal of this study was not to define measures to be taken to achieve the best possible interobserver agreement but to assess the reliability of the CTF nomenclature and to compare it with that of the Nordic nomenclature in conditions as close as possible to routine clinical practice.

Reports on disk contour abnormalities should be standardized, as far as possible. However, the fact that reports are standardized does not necessarily mean that they provide clinically relevant information.[2-4,29] In fact, correlation between clinical and radiologic findings should always be prioritized.[2]

Current recommendations for grading degenerative spine disease propose the use of scales with 3–5 grades, starting with the "not degenerated" state.[30] Both the CTF and the Nordic nomenclatures follow this recommendation. In general, it is assumed that diagnostic scales with more categories lead to lower agreement, higher sensitivity and specificity, and narrower CIs.[31] Nevertheless, in a previous study, the degree of interobserver agreement between 2 experienced readers decreased when they were forced to reduce the number of diagnostic categories from 3 (normal, bulge, herniation) to 2 (no herniation versus herniation).[7] This finding is consistent with results from the current study, in which only 2 categories were analyzed (normal versus abnormal), and interobserver agreement varied noticeably, depending on whether bulging was categorized as normal or abnormal.

The κ statistic is hampered by low and high prevalences.[19] For this reason, in this study, κ was calculated only for findings with a prevalence between 10% and 90%. The interpretation of κ values may be seen as challenging because there is not a clear threshold indicating when a κ value becomes inconsistent with high-quality clinical care.[18] Furthermore, it is difficult to compare κ values across studies in which categories or the prevalence of findings is different. Nevertheless, the κ value probably remains the best available method of measuring concordance, which is in addition to that explained by chance.

This study has some other limitations. Reasons for discrepancies were not analyzed. However, these discrepancies were not the goal of this study, and this approach has been shown not to resolve disagreement.[26] No attempt was made to evaluate reader consistency beyond the terminology of disk contour (ie, nerve root compression), and readers were not asked to situate those findings in the spinal canal or neural structures. However, this was not our objective of this study; and though such descriptions are a key component of any radiologic report, the best interobserver agreement on such features, from readers working at the same institution, yielded a κ score of 0.67.[11] The prevalence of findings did not permit statistically sound comparisons in reproducibility beyond the L4-L5 and L5-S1 levels. The need to collapse categories might have been reduced with a larger sample. However, this would only be possible if the prevalence of findings had been between 10% and 90%, which is difficult to guarantee in a clinical setting.[22]

Implications for Care

In clinical practice, although reports from the same radiologist are reasonably consistent, only moderate agreement among radiologists can realistically be expected in the interpretation of lumbar disk contour. The degree of agreement is similar regardless of the nomenclature used. Standardization of nomenclature increases self-confidence in reporting, but the correlation between clinical and radiologic findings should always be prioritized.

Conclusions

Results from this study show that in conditions that are as close as possible to clinical practice, the interobserver agreement in the interpretation of intervertebral herniation and disk contour on lumbar MR imaging is, at best, moderate, irrespective of whether the CTF or Nordic nomenclatures are used. The latter increases self-confidence in an individual observer's report but is less clear regarding the classification of disks as normal versus bulged.

References

1. Deyo RA, Mirza SK, Turner JA, et al. **Overtreating chronic back pain: time to back off?** *J Am Board Fam Med* 2009;22:62–68
2. Milette PC. **The proper terminology for reporting lumbar intervertebral disk disorders.** *AJNR Am J Neuroradiol* 1997;18:1859–66
3. Chou R, Fu R, Carrino JA, et al. **Imaging strategies for low-back pain: systematic review and meta-analysis.** *Lancet* 2009;373:463–72
4. Bradley WG Jr, for the Expert Panel on Neurologic Imaging. **Low back pain**. *AJNR Am J Neuroradiol* 2007;28:990–92
5. Jarvik JG, Deyo RA. **Diagnostic evaluation of low back pain with emphasis on imaging.** *Ann Intern Med* 2002;137:586–97
6. Carrino JA, Lurie JD, Tosteson AN, et al. **Lumbar spine: reliability of MR imaging findings.** *Radiology* 2009;250:161–70
7. Brant-Zawadzki MN, Jensen MC, Obuchowski N, et al. **Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities: a comparison of two nomenclatures.** *Spine (Phila Pa 1976)* 1995;20:1257–63, discussion 1264
8. Jarvik JG, Haynor DR, Koepsell TD, et al. **Interreader reliability for a new classification of lumbar disk disease.** *Acad Radiol* 1996;3:537–44
9. Arana E, Royuela A, Kovacs FM, et al. **Agreement in the interpretation of magnetic resonance images of the lumbar spine using the Nordic Modic Consensus Group Classification form.** *Radiology* 2010;254:809–17
10. Kovacs FM, Royuela A, Jensen TS, et al. **Agreement in the interpretation of magnetic resonance images of the lumbar spine.** *Acta Radiol* 2009;50:497–506
11. Weishaupt D, Zanetti M, Hodler J, et al. **MR imaging of the lumbar spine: prevalence of intervertebral disk extrusion and sequestration, nerve root compression, end plate abnormalities, and osteoarthritis of the facet joints in asymptomatic volunteers.** *Radiology* 1998;209:661–66
12. Lurie JD, Tosteson ANA, Tosteson TD, et al. **Reliability of magnetic resonance imaging readings for lumbar disc herniation in the Spine Patient Outcomes Research Trial (SPORT).** *Spine* 2008;33:991–98
13. Nachemson A. **Back pain: delimiting the problem in the next millennium.** *Int J Law Psychiatry* 1999;22:473–90

14. Fardon DF, Milette PC, for the Combined Task Forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology. **Nomenclature and classification of lumbar disc pathology: recommendations of the combined task forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology**. *Spine (Phila Pa 1976)* 2001;26:E93–113
15. Solgaard Sorensen J, Kjaer P, Jensen ST, et al. **Low-field magnetic resonance imaging of the lumbar spine: reliability of qualitative evaluation of disc and muscle parameters.** *Acta Radiol* 2006;47:947–53
16. Ross JS. **Babel 2.0**. *Radiology* 2010;254:640–41
17. Jensen TS, Sorensen JS, Kjaer P. **Intra- and interobserver reproducibility of vertebral endplate signal (modic) changes in the lumbar spine: the Nordic Modic Consensus Group classification.** *Acta Radiol* 2007;48:748–54
18. Landis JR, Koch GG. **The measurement of observer agreement for categorical data.** *Biometrics* 1977;33:159–74
19. Feinstein AR, Cicchetti DV. **High agreement but low kappa. I. The problems of two paradoxes.** *J Clin Epidemiol* 1990;43:543–49
20. Lipsitz SR, Parzen M, Fitzmaurize GM, et al. **A two-stage logistic regression model for analyzing inter-rater agreement.** *Psychometrika* 2003;68:289–98
21. Hardin JW, Hilbe JM. *Generalized Estimating Equations.* Boca Raton, Florida: Chapman & Hall; 2003
22. Jarvik JG, Deyo RA. **Moderate versus mediocre: the reliability of spine MR data interpretations.** *Radiology* 2009;250:15–17
23. Gur D, Bandos AI, Fuhrman CR, et al. **The prevalence effect in a laboratory environment: changing the confidence ratings.** *Acad Radiol* 2007;14:49–53
24. Ng CS, Palmer CR. **Analysis of diagnostic confidence and diagnostic accuracy: a unified framework.** *Br J Radiol* 2007;80:152–60
25. Carragee EJ, Alamin TF, Miller JL, et al. **Discographic, MRI and psychosocial determinants of low back pain disability and remission: a prospective study in subjects with benign persistent back pain.** *Spine J* 2005;5:24–35
26. van Rijn JC, Klemetso 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
27. Brorson S, Hrobjartsson A. **Training improves agreement among doctors using the Neer system for proximal humeral fractures in a systematic review.** *J Clin Epidemiol* 2008;61:7–16
28. Lurie JD, Tosteson AN, Tosteson TD, et al. **Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis.** *Spine* 2008;33:1605–10
29. Gillan MG, Gilbert FJ, Andrew JE, et al. **Influence of imaging on clinical decision making in the treatment of lower back pain.** *Radiology* 2001;220:393–99
30. Kettler A, Wilke HJ. **Review of existing grading systems for cervical or lumbar disc and facet joint degeneration.** *Eur Spine J* 2006;15:705–18
31. Bailey IL, Bullimore MA, Raasch TW, et al. **Clinical grading and the effects of scaling.** *Invest Ophthalmol Vis Sci* 1991;32:422–32