Lumbar Spine: Agreement in the Interpretation of 1.5-T MR Images by Using the Nordic Modic Consensus Group Classification Form[1]

Estanislao Arana, MD, PhD, MHA
Ana Royuela, MSc
Francisco M. Kovacs, MD, PhD
Ana Estremera, MD
Helena Sarasíbar, MD
Guillermo Amengual, MD
Isabel Galarraga, MD
Carmen Martínez, MD
Alfonso Muriel, MSc
Víctor Abraira, PhD
María Teresa Gil del Real, MPH
Javier Zamora, PhD
Carlos Campillo, MD

Purpose: To evaluate intra- and interobserver agreement for the interpretation of lumbar 1.5-T magnetic resonance (MR) images in a community setting.

Materials and Methods: The study design was approved by the Institutional Review Board of the Ramón y Cajal Hospital. According to Spanish law, for this type of study, no informed consent was necessary. Five radiologists from three hospitals twice interpreted lumbar MR examination results in 53 patients with low back pain, with at least a 14-day interval between assessments. Radiologists were unaware of the clinical and demographic characteristics of the patients and of their colleagues' assessments. At the second assessment, they were unaware of the results of the first assessment. Reports on Modic changes, osteophytes, Schmorl nodes, diffuse defects, disk degeneration, annular tears (high-signal-intensity zones), disk contour, spondylolisthesis, and spinal stenosis were collected by using the Spanish version of the Nordic Modic Consensus Group classification. The κ statistic was used to assess intra- and interobserver agreement for findings with a prevalence of 10% or greater and 90% or lower. κ was 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), or poor (<0.00).

Results: Endplate erosions and spondylolisthesis were observed in less than 10% of images. Intraobserver reliability was almost perfect for spinal stenosis; substantial for Modic changes, Schmorl nodes, disk degeneration, annular tears, and disk contour; and moderate for osteophytes. Interobserver reliability was moderate for Modic changes, Schmorl nodes, disk degeneration, annular tears, and disk contour; fair for osteophytes; and poor for spinal stenosis.

Conclusion: In conditions close to those of clinical practice, there was only moderate interobserver agreement in the reporting of findings at 1.5-T lumbar MR imaging.

Supplemental material: *http://radiology.rsna.org/lookup/suppl/doi:10.1148/radiol.09090706/-/DC1*

[1] From the Department of Radiology, Hospital Quirón, Valencia, Spain (E.A.); Spanish Back Pain Research Network (E.A., A.R., F.M.K., A.E., H.S., G.A., I.G., C.M., A.M., V.A., M.T.G.d.R., J.Z., C.C.) and Scientific Department (F.M.K., M.T.G.d.R.), Fundación Kovacs, Paseo de Mallorca 36, 07012 Palma de Mallorca, Spain; CIBER Epidemiología y Salud Pública (A.R., A.M., V.A., J.Z.) and Clinical Biostatistics Unit (A.R., A.M., V.A., J.Z.), Hospital Ramón y Cajal, Madrid, Spain; Department of Radiology, Hospital Son Llàtzer, Palma de Mallorca, Spain (A.E., H.S., G.A., C.M.); Department of Radiology, Hospital de Manacor, Manacor, Mallorca, Spain (I.G.); and Ib-Salut, Palma de Mallorca, Spain (C.C.) Received May 14, 2009; revision requested July 2; revision received August 17; accepted August 28; final version accepted September 16. Supported by the Kovacs Foundation, Palma de Mallorca, Spain. **Address correspondence to** F.M.K. (e-mail: *fmkovacs@kovacs.org*).

Magnetic resonance (MR) imaging of the lumbar spine is a frequently performed procedure (1). Like many imaging tests, the reliability of interpretation of its findings is to some degree taken for granted. However, several studies (2–22) have been undertaken to assess the concordance of different physicians in reporting findings such as disk degeneration, Modic changes, annular tears, disk bulges, protrusions and herniations, and spinal stenosis at MR imaging. In general, the concordance found in those studies ranged from moderate to excellent, depending on the MR imaging finding that was being evaluated. However, these studies may have resulted in overestimation of that concordance, as it was often between only two or three highly specialized readers who, in most cases, worked together in a research setting. This may have led to an informal agreement in their diagnostic criteria (3–16). Only seven studies, focusing on Modic changes (7,9,18,19), disk degeneration (9,18–20), spondylolisthesis (9,18), and spinal stenosis (15,19,21), assessed the concordance of more than three observers, and most of these studies still included only physicians who worked together in research settings. Therefore, there is a potential need to assess the reliability of these interpretations in situations more like clinical practice (22,23).

Advances in Knowledge

- In conditions similar to those of clinical practice, intraobserver reliability of the interpretation of 1.5-T lumbar MR imaging findings among five radiologists was almost perfect for spinal stenosis; substantial for Modic changes, Schmorl nodes, disk degeneration, annular tears, and disk contour; and moderate for osteophytes.
- Interobserver reliability was moderate for Modic changes, Schmorl nodes, disk degeneration, annular tears, and disk contour and was fair for osteophytes.

Thus, the purpose of this study was to assess the intra- and interobserver concordance in the interpretation of 1.5-T lumbar spine MR images in patients seen in clinical practice among radiologists who had not previously agreed on diagnostic criteria and who work clinically in different locations.

Materials and Methods

This study was funded by the Kovacs Foundation, a nonprofit Spanish research institution with its own funding and no links to the health industry.

Study Population

The study was approved by the Institutional Review Board of the Ramón y Cajal Hospital on April 24, 2009.

Five practicing general radiologists (E.A., A.E., H.S., G.A., and I.G., with 12, 8, 7, 10, and 1 year of experience, respectively, in interpreting spine images), working in three general hospitals located in two different geographic regions in Spain, participated in this study. Their postresidency experience as radiologists ranged from 12 to 18 years. They were trained in different institutions, and none had formal fellowship training.

Two of those radiologists (E.A., A.E.) working in hospitals in different cities saw images of patients who had undergone 1.5-T MR imaging for low back pain and/or sciatica. They randomly selected images from 53 of these patients. According to Spanish law for a study such as this one, in which the images of the patients had already been obtained and were anonymized, informed consent is not required.

Exclusion criteria were as follows: previous spinal surgery, pregnancy, cauda equina syndrome, scoliosis with a more than 15° curvature, vertebral fractures, inflammatory spondyloarthropathy, spinal infection, or tumor. Exclusions were as follows: seven patients because of previous spinal surgery, five patients because of scoliosis, and three patients because of metastatic cancer.

Twenty-eight female and 25 male patients were studied (mean age, 48 years ± 13.3 [standard deviation]). Mean age for men was 46.3 years ± 13.7, and mean age for women was 50.3 years ± 12.9 (P = .274).

Implication for Patient 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 1.5-T MR examination findings.

MR Imaging

Patients who were selected had undergone MR imaging performed with one of two 1.5-T systems (Genesis Signa, GE Medical Systems, Milwaukee, Wis; ACS Intera NT Gyroscan, Philips Medical Systems, Eindhoven, the Netherlands) employing phased-array multicoils. All patients were studied in the supine position with a fixed imaging protocol (Table 1).

All images were masked as to name, sex, and age and were distributed to all radiologists participating in this study.

Variables

Radiologists were asked to report their findings by using the Spanish version of the Nordic Modic Consensus Group classification (6,19), in which the following variables were recorded separately for all the lumbar segments (from L1-L2 to L5-S1).

Published online before print
10.1148/radiol.09090706

Radiology 2010; 254:809–817

Author contributions:
Guarantors of integrity of entire study, F.M.K., H.S., M.T.G.d.R., C.C.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, E.A., F.M.K., M.T.G.d.R., C.C.; clinical studies, E.A., A.E., H.S., G.A., I.G., C.M., C.C.; statistical analysis, A.R., A.M., V.A., J.Z., C.C.; and manuscript editing, E.A., A.R., F.M.K., A.E., G.A., I.G., C.M., A.M., V.A., M.T.G.d.R., J.Z., C.C.

Authors stated no financial relationship to disclose.

See also the editorial by Ross in this issue.

Table 1

Sequences for 53 MR Imaging Examinations

Pulse Sequence	Repetition Time (msec)/ Echo Time (msec)	Field of View (mm)	Matrix	No. of Signals Acquired	Section Thickness (mm)	Flip Angle (degrees)	Intersection Gap (mm)	Echo Train Length
Localizer gradient echo	30/10	400	128 × 128	1	10	50	...	...
Sagittal T1-weighted spin echo	440–550/14–20	270	156–307 × 192–512	2	4	...	0.4–1.3	...
Sagittal T2-weighted turbo spin echo	3300–2896/102.9–120	270	156–307 × 192–512	2	4	...	0.4–1.3	12
Axial T2-weighted turbo spin echo (parallel to disk spaces)	3040–2896/103–120	180	224–190 × 256–512	3	4	...	0.4	5

For variables related to Modic changes (24) (predominant and nonpredominant), presence (no changes or type 1, 2, or 3), location (the upper or lower vertebral endplate), maximum height (craniocaudally) affected by Modic changes (only the endplate, <25% of the vertebral body, between 25% and 50% of the vertebral body, >50% of the vertebral body), maximum volume (craniocaudally) affected by Modic changes (only the endplate, <25% of the vertebral body, between 25% and 50% of the vertebral body, >50% of the vertebral body), endplate extension affected by Modic changes in the anteroposterior axis (<25% of the endplate, between 25% and 50% of the endplate, >50% of the endplate), and maximum endplate area affected by Modic changes (<25% of the endplate area, between 25% and 50% of the endplate area, between 51% and 75% of the endplate area, >75% of the endplate area) were recorded. On those images in which different types of Modic changes were observed, those variables were assessed separately for predominant (ie, most widely observed in that particular level) and nonpredominant changes. At the analysis phase, type of Modic change was dichotomized into no changes versus change (type 1, 2, or 3) categories, and maximum affected height, maximum affected volume (craniocaudally), affected endplate area, and affected endplate extension (anteroposterior) were dichotomized into no changes versus all the other categories.

For variables related to other bone changes, osteophytes (yes or no and location), Schmorl nodes, and endplate erosions were noted. They include localized endplate erosions (yes or no and location), and whether those were located beside concomitant Modic changes (yes or no), and irregular endplate—as in Scheuermann disease (yes or no). Localized defects were defined as "sharp" indentations or discontinuity of the cortical bone. Irregular endplates were defined as endplates that were intact but irregular (6).

For variables related to disk changes, annular tears (fissure or high-signal-intensity zones) in the annulus fibrosus assessed on T2-weighted images (yes or no), signs of disk degeneration (grade according to Pfirrmann classification [11], Table E1 [online]), and disk contour (normal, bulging disk, protrusion [focal or broad-based], or hernia [extrusion or sequestration]) were noted. At the analysis phase, disk contour was dichotomized in normal versus abnormal (bulging disk, protrusion, or extrusion) groups, and disk degeneration was dichotomized into grades I, II, and III versus IV and V.

For variables related to other findings, presence of spinal stenosis, defined as any type of acquired narrowing of the spinal canal (25) (yes or no), and spondylolisthesis (none or grade according to Meyerding classification, although it was dichotomized at the analysis phase into no spondylolisthesis vs grades I–IV) (11,26) were recorded.

Assessment and Data Collection

All MR images were presented on compact discs created by using imaging software (K-PACS, 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 five radiologists were unaware of any demographic and clinical data from the patients from whom the images had been obtained. They were asked to report their findings on a structured form with close-ended responses for each variable (the Spanish version of the Nordic Modic Consensus Group classification), no matter what their opinion was about the potential clinical relevance of those findings (6,19).

Besides the above, they were asked to act as they usually do in their routine clinical practice. No attempt was made to homogenize their diagnostic criteria, and they received no instructions regarding the interpretation of images. They assessed the MR images alone and had no access to reports from their colleagues.

To assess intraobserver reliability, the five radiologists were asked to reevaluate the same MR images at least 14 days after the forms with their first interpretation had been collected. Radiologists were unaware that the images they assessed in the second round were the same as those they had interpreted in the first one. The images were presented in a different order, and radiologists had no access to their previous

Table 2

Predominant Findings in 53 MR Imaging Examination Reports from Five Radiologists

A: Findings Related to Modic and Bone Changes

	Segment L1		Segment L2		Segment L3		Segment L4		Segment L5		
MR Imaging Variable	U	L	U	L	U	L	U	L	U	L	Segment S1*
Modic change											
Presence											
No changes	229	235	222	232	219	218	189	196	188	143	139
Type 1	2	1	2	2	3	4	5	8	3	7	9
Type 2	34	29	41	31	43	43	71	61	73	115	116
Type 3	0	0	0	0	0	0	0	0	1	0	1
Maximum affected height											
No changes	229	230	222	232	219	216	190	195	190	139	138
Only the endplate	16	22	20	20	29	28	49	33	33	55	68
<25% of the vertebral body	12	10	17	10	14	10	18	21	22	37	44
25%–50% of the vertebral body	8	3	6	3	3	8	8	12	17	26	14
>50% of the vertebral body	0	0	0	0	0	3	0	4	3	8	1
Maximum affected volume											
No changes	229	230	222	232	219	216	190	195	190	139	138
Only the endplate	16	22	20	20	29	30	51	32	35	58	71
<25% of the vertebral body	14	11	19	10	15	13	18	26	29	51	44
25%–50% of the vertebral body	6	2	4	3	2	5	6	11	10	16	12
>50% of the vertebral body	0	0	0	0	0	1	0	1	1	1	0
Affected endplate area											
No changes	229	230	222	232	219	216	190	195	190	139	138
<25% of endplate area	25	28	29	25	37	33	62	48	48	76	83
25%–50% of endplate area	11	5	13	7	8	15	13	13	18	33	27
>50% of endplate area	0	2	1	1	1	1	0	9	9	17	17
Maximum affected extension (anteroposterior)											
No changes	229	230	222	232	219	216	190	195	190	141	141
<25% of endplate	19	24	26	21	30	28	49	38	42	56	57
25%–50% of the endplate	12	7	9	5	14	13	14	9	12	21	15
51%–75% of the endplate	5	4	7	5	2	4	7	8	8	17	25
>75% of the endplate	0	0	1	2	0	4	5	15	13	30	27
Osteophyte											
No. of osteophytes	42	69	79	70	70	84	105	114	131	120	95
Located beside Modic change (yes)	12	23	28	16	30	29	50	38	48	72	64
Schmorl node											
No. of Schmorl nodes (yes)	19	29	38	29	35	18	44	26	23	17	10
Located beside Modic change (yes)	5	3	13	1	9	4	18	11	14	12	6
Endplate erosion											
No. of endplate erosions (irregular endplate) (yes)	6	8	8	4	5	6	10	17	13	29	22
Located beside Modic change (yes)	0	2	3	1	1	2	5	13	12	20	15

B: Findings Related to Disk Changes and Other Findings

MR Imaging Variable	L1-L2 Level	L2-L3 Level	L3-L4 Level	L4-L5 Level	L5-S1 Level
Disk degeneration (Pfirrmann grade)†					
I	20	18	17	19	17
II	94	80	65	27	50
III	106	114	106	102	84

Table 2 (continues)

Table 2 (continued)

Predominant Findings in 53 MR Imaging Examination Reports from Five Radiologists

B: Findings Related to Disk Changes and Other Findings					
MR Imaging Variable	L1-L2 Level	L2-L3 Level	L3-L4 Level	L4-L5 Level	L5-S1 Level
IV	32	47	71	103	74
V	13	6	6	14	40
Annular tears (yes)	2	9	25	82	74
Disk contour					
Normal	217	204	160	84	104
Bulging disk	45	46	97	137	98
Protrusion	3	11	8	39	60
Extrusion (contained or uncontained)	0	4	0	5	3
Spondylolisthesis					
None	265	264	264	254	257
I	0	1	1	11	8
II	0	0	0	0	0
III	0	0	0	0	0
IV	0	0	0	0	0
Spinal stenosis (yes)	13	17	24	48	42

Note.—Data are numbers of findings. The number of reports was 265 (53 images interpreted by five radiologists). κ Values were not calculated for findings reported in 27 or fewer or 238 or more of those reports. L = lower endplate, U = upper endplate.

* Findings for segment S1 are for the upper endplate.

† For the purpose of this investigation, grades I–III (ie, healthy adolescent [grade I], healthy adult [grade II], and early degeneration [grade III]) were grouped together, as were the more advanced grades of degeneration (ie, grades IV and V).

reports or to the current or previous reports of their colleagues.

All reports were entered in the database at a centralized coordination office. Entry of data was done independently by two administrative assistants who verified that the data they were entering coincided with the information on the forms.

Statistical Analysis

To assess intra- and interobserver reliability, ratings from each observer were cross-tabulated, and agreement was measured by using the κ statistic, which was 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 (27).

The κ statistic is affected 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 (14). Therefore, at the design phase, it was decided that κ values would be calculated only for findings reported in more than 10% and in less than 90% of reports. Because five radiologists interpreted 53 images (total, 265 reports), κ values were not calculated for findings identified in 27 or fewer or in 238 or more of those reports.

To make it possible to analyze results by using the κ statistic, the following strategy was used: *(a)* All the response levels were dichotomized into two categories (normal vs abnormal). *(b)* Findings at each level (ie, L1, L1-L2, etc) were listed, and those for which there was a prevalence between 10% and 90% were identified. *(c)* κ was calculated following the two-step approximation described by Lipsitz et al (28). This approximation basically consists of estimating the expected and observed probabilities by means of logistic regression. In this case, the regression model included age, sex, and their interaction, and generalized estimating equation (29) models were used. The structure was self-regressive correlation to establish the existent correlation between the different vertebral levels on an image. Means and 25th and 75th percentiles of those κ values were determined.

Statistical packages (STATA IC/10.0 for Windows, Stata Statistical Software, College Station, Tex; SPSS, version 16.0, SPSS, Chicago, Ill) were used for data analysis.

Results

Tables 2 and 3 show the findings reported by the five radiologists at 53 MR imaging examinations, and those findings are shown in detail in Tables E2–E6 (online).

Most findings related to Modic changes, disk degeneration, bulging disk or protrusion, osteophytes, spondylolisthesis, and spinal stenosis were found at the L4-L5 and L5-S1 levels, while most Schmorl nodes were reported at the L1-L2 and L2-L3 levels (Tables 2 and 3).

The low number of nonpredominant Modic changes, endplate erosions, and spondylolisthesis made these unsuitable for κ calculation.

For the same reason, κ values for variables related to Schmorl nodes and high-signal-intensity zones could be calculated only for some levels. On the

Table 3

Nonpredominant Findings in 53 MR Imaging Examination Reports from Five Radiologists

	Segment L1		Segment L2		Segment L3		Segment L4		Segment L5		
MR Imaging Variable (Modic Change)	U	L	U	L	U	L	U	L	U	L	Segment S1*
Presence											
No changes	264	259	258	263	262	261	263	258	252	258	257
Type 1	0	3	4	2	2	2	2	5	6	5	2
Type 2	1	1	2	0	1	1	0	2	6	2	4
Type 3	0	2	1	0	0	1	0	0	1	0	2
Affected maximum height											
No changes	264	260	258	263	262	261	263	259	253	259	257
Only the endplate	0	2	1	1	2	1	1	4	6	2	6
<25% of the vertebral body	1	2	5	1	1	2	0	2	4	3	2
25%–50% of the vertebral body	0	1	1	0	0	0	1	0	2	1	0
>50% of the vertebral body	0	0	0	0	0	1	0	0	0	0	0
Maximum affected volume											
No changes	264	259	258	263	262	261	263	259	253	259	257
Only the endplate	0	3	1	1	2	1	1	4	7	2	6
<25% of the vertebral body	1	3	6	1	1	2	1	2	4	4	2
25%–50% of the vertebral body	0	0	0	0	0	1	0	0	1	0	0
>50% of the vertebral body	0	0	0	0	0	0	0	0	0	0	0
Affected endplate area											
No changes	264	259	258	263	262	261	263	259	253	259	257
<25% of endplate area	0	5	5	2	3	3	1	5	9	5	6
25%–50% of endplate area	1	0	2	0	0	0	0	1	2	1	2
>50% of endplate area	0	1	0	0	0	1	1	0	1	0	0
Maximum affected extension (anteroposterior)											
No changes	264	259	258	263	262	261	263	259	253	259	257
<25% of the endplate	0	4	4	1	2	3	1	5	7	3	3
25%–50% of the endplate	1	1	2	1	0	0	0	1	2	3	2
51%–75% of the endplate	0	1	1	0	0	0	0	0	2	0	3
>75% of the endplate	0	0	0	0	1	1	1	0	1	0	0

Note.—Data are numbers of findings. The number of reports was 265 (53 images interpreted by five radiologists). κ Values were not calculated for findings reported in 27 or fewer or 238 or more of those reports. L = lower endplate, U = upper endplate.

* Findings for segment S1 are for the upper endplate.

contrary, variables related to predominant Modic changes, osteophytes, signs of disk degeneration, and disk contour could be calculated for all levels (Tables 2 and 3).

The intraobserver agreement was almost perfect for variables related to spinal stenosis; substantial for variables related to Modic changes, Schmorl nodes, disk degeneration, annular tears, and disk contour; and moderate for variables related to osteophytes (Table 4).

The interobserver agreement was moderate for Modic changes, Schmorl nodes, disk degeneration, annular tears, and disk contour and was fair for osteophytes. It was not possible to calculate agreement for spinal stenosis, because convergence was not achieved at the first generalized estimating equation analysis (Table 5).

Intraobserver agreement was always higher than interobserver agreement (Tables 4 and 5).

Discussion

The Nordic Modic Consensus Group classification is a structured form to gather findings on MR images which has proved to be reliable for 0.23-T MR images in previous studies performed with the same images in two different countries (6,19). Dissemination of this classification system has been encouraged (24).

This use of the Nordic Modic classification to assess agreement in the interpretation of 1.5-T lumbar MR images yielded results generally comparable to those obtained for 0.23 T. Except for annular tears (high-signal-intensity zones), interobserver agreement was only moderate for most variables and was slight for osteophytes, and there was rare agreement for spinal stenosis (6,19).

For all MR imaging findings, intraobserver agreement was higher than interobserver agreement. This may be

Table 4

Intraobserver Agreement in the Interpretation of Lumbar MR Images by Five Radiologists

MR Imaging Variable	No. of Levels Evaluated	No. of Levels Analyzed*	Intraobserver Agreement†
Modic change	11‡	10	0.724 (0.691, 0.759)
Affected maximum height (normal vs abnormal)	11‡	10	0.724 (0.690, 0.757)
Maximum volume (craniocaudal) (normal vs abnormal)	11‡	10	0.728 (0.694, 0.762)
Affected endplate area (normal vs abnormal)	11‡	10	0.725 (0.688, 0.758)
Maximum affected extension (anteroposterior) (normal vs abnormal)	11‡	10	0.720 (0.681, 0.755)
Osteophyte			
Osteophytes (yes or no)	11‡	11	0.513 (0.495, 0.521)
Located beside Modic change (yes or no)	11‡	5	0.648 (0.610, 0.688)
Schmorl nodes (yes or no)	11‡	5	0.754 (0.714, 0.835)
Endplate erosions (yes or no)	11‡	0	Prevalence too low to calculate κ values
Disk degeneration (Pfirrmann grade) (normal or abnormal)	5§	5	0.689 (0.651, 0.722)
Annular tears (yes or no)	5§	2	0.686 (0.662, 0.727)
Disk contour (normal vs abnormal)	5§	5	0.728 (0.664, 0.781)
Spondylolisthesis (normal vs abnormal)	5§	0	Prevalence too low to calculate κ values
Spinal stenosis (yes or no)	5§	2	0.870 (0.847, 0.896)

* κ Analyses.

† Data are mean κ values, with 25th and 75th percentiles in parentheses.

‡ The maximum number of levels that could have been evaluated was 11 (superior and inferior endplates for L1-L2 to L5-S1, plus S1).

§ The maximum number of levels that could have been evaluated was five (spaces from L1-L2 to L5-S1).

interpreted as suggesting that clinicians can expect reasonably consistent reports from a given radiologist but should be aware that those reports would not necessarily be consistent with those from other radiologists. This might encourage clinicians to preferentially refer their patients to the same radiologist, whose criteria and style of reporting they feel more confident with.

A recent study (18) in a different setting has assessed the agreement on the reporting of Modic changes, disk degeneration, high-signal-intensity zones (annular tears), spondylolisthesis, and facet arthropathy at 1.5-T MR imaging. Although methods similar to those implemented in the current study were used, that study did not measure agreement on disk contour descriptors, and the Nordic Modic Consensus Group classification form was not used. That study assessed MR images obtained during the Spine Patient Outcomes Research Trial, a high-quality trial with a relatively homogeneous sample that included only patients who were candidates for surgery. MR images were assessed by four highest-level expert readers (radiologists and orthopedic surgeons, both with fellowship training, with more than 12 years of experience reading spine MR images) by using diagnostic criteria that had been defined in a consensus meeting and were included in an available handbook (18). On the contrary, in the current study, images were obtained in patients referred for lumbar MR imaging in routine clinical practice who were not necessarily candidates for surgery. The images were assessed in the radiology departments of three different hospitals by five radiologists who were unaware of patients' clinical features and had not agreed on any diagnostic criteria. In spite of those differences, results from both studies were consistent with regard to the agreement found for Modic changes. Interobserver agreement in the current study was worse for disk degeneration and slightly better for annular tears, and intraobserver agreement was similar for all common variables.

Interobserver agreement for disk contour in this study (κ = 0.546) was higher than that observed for radiologists' and clinicians' agreement when interpreting herniation morphology in the Spine Patient Outcomes Research Trial (κ = 0.24), in which a semistandardized nomenclature was used in the clinical setting (20).

In this study, interobserver agreement for spinal stenosis was poor. Other studies (15) in which, as in this one, readers were not given any instructions or diagnostic criteria, also showed low interobserver agreement. Conversely, studies (18,21) in which specific ad hoc training and clear-cut instructions were provided to the reader showed better interobserver agreement. A three-category classification for spinal stenosis (mild, moderate, and severe) has shown to be even more reliable than quantitative measures (21). In this prior study, intraobserver agreement for spinal stenosis was almost perfect, suggesting that the use of a previously agreed on diagnostic criteria would have improved interobserver agreement.

The interobserver agreement for the extension of Modic changes observed in this study was similar to that in the study deriving from the Spine Patient Outcomes Research Trial (18) and slightly worse than that in the study in which the Nordic Modic classification form was developed by using 0.23-T MR images (6). That latter study did not use images of patients from clinical practice, but from a representative sample of the general Danish population at age 40; the readers knew that the population presented a 69% prevalence of Modic changes, and images were interpreted at a research department in which consensus training meetings were implemented among

Table 5

Interobserver Agreement in the Interpretation of Lumbar MR Images by Five Radiologists

MR Imaging Variable	No. of Levels Evaluated	No. of Levels Analyzed*	Interobserver Agreement†
Modic change			
Presence	11‡	11	0.526 (0.456, 0.593)
Affected maximum height (normal vs abnormal)	11‡	11	0.523 (0.458, 0.589)
Maximum affected volume (craniocaudal) (normal vs abnormal)	11‡	11	0.523 (0.458, 0.589)
Affected endplate area (normal vs abnormal)	11‡	11	0.523 (0.458, 0.589)
Maximum affected extension (anteroposterior) (normal vs abnormal)	11‡	11	0.511 (0.440, 0.583)
Osteophyte			
Osteophytes (yes or no)	11‡	11	0.364 (0.341, 0.375)
Located beside Modic change (yes or no)	11‡	8	0.560 (0.501, 0.633)
Schmorl nodes (yes or no)	11‡	5	0.539 (0.454, 0.630)
Endplate erosions (yes or no)	11‡	0	Prevalence too low to calculate κ values
Disk degeneration (Pfirrmann grade) (normal or abnormal)	5§	5	0.491 (0.430, 0.553)
Annular tears (yes or no)	5§	2	0.599 (0.557, 0.632)
Disk contour (normal vs abnormal)	5§	5	0.546 (0.494, 0.587)
Spondylolisthesis (normal vs abnormal)	5§	0	Prevalence too low to calculate κ values
Spinal stenosis (yes or no)	5§	2	Convergence not achieved

* κ Analyses.

† Data are mean κ values, with 25th and 75th percentiles in parentheses.

‡ The maximum number of levels that could have been evaluated was 11 (superior and inferior endplates for L1-L2 to L5-S1, plus S1).

§ The maximum number of levels that could have been evaluated was five (spaces from L1-L2 to L5-S1).

the three specialists in charge of assessing the images (6,14). Those features may contribute to the slight differences in the results from that study.

This study had some potential limitations. The variability in image acquisition methods introduced some heterogeneity; however, this variability corresponds to the one commonly encountered in routine clinical practice, and all the radiologists interpreted the same images. The heterogeneity of patients in whom MR imaging findings were studied was also likely to influence the degree of agreement in the interpretation of the images (30). On the one hand, to retain some homogeneity, patients with scoliosis with a greater than 15° curvature, known systemic diseases (eg, inflammatory spondyloarthropathy or cancer), or cauda equina syndrome and those who had undergone previous back surgery were excluded. On the other hand, our study sample was as heterogeneous as is to be expected in routine clinical conditions; some patients may have been clinically eligible for surgery for spinal stenosis or disk herniation, whereas others may have had common low back pain with no radicular pain or MR imaging may have been ordered only to exclude potential systemic diseases. Even in prior studies (18) with highly selected populations of surgical candidates recruited in a high-quality trial such as the Spine Patient Outcomes Research Trial, the interobserver agreement for the interpretation of MR imaging findings was only moderate and similar to that of this study. Because it is easier to analyze agreement when a structured form is used to report the findings on MR images, the Nordic Modic classification form was used in this study. In this study, only the presence or absence of stenosis was analyzed, rather than its severity and location. However, absolute values and correlation between stenosis measurements and symptoms appear to be lacking (31). This was another possible limitation.

The interpretation of κ values may be seen as challenging, because there is no clear threshold indicating when a κ value becomes inconsistent with high-quality clinical care (14,32). Furthermore, it is difficult to compare κ values across studies in which categories or the prevalence of findings are different. Last, κ values decrease when the prevalence of the finding is very high or very low, even when the observed agreement remains unchanged (32). Nevertheless, the κ value probably remains the best available method of measuring concordance, in addition to that explained by chance.

This study was designed to be performed in conditions similar to routine practice. Several measures might be taken to improve interobserver agreement in the interpretation of lumbar MR imaging findings, such as agreeing on diagnostic criteria, further improving reader training, and using available online examples and linking them to standardized nomenclature (22,30). However, the feasibility of implementing those measures in routine clinical practice is unknown, and interobserver agreement in studies in which some of those measures were implemented was still moderate and, for most variables, not radically different from the one in the current study (18). This shows that, even though those measures may improve interobserver agreement, they are not without their own difficulties (22). Taken together, results from these studies suggest that, in practice, only moderate agreement can realistically be expected in routine practice.

References

1. Airaksinen O, Brox JI, Cedraschi C, et al. Chapter 4 European guidelines for the management of chronic nonspecific low back pain. Eur Spine J 2006;15(suppl 2):S192–S300.

2. Hollingworth W, Gray DT, Martin BI, Sullivan SD, Deyo RA, Jarvik JG. Rapid magnetic resonance imaging for diagnosing cancer-related low back pain. J Gen Intern Med 2003;18(4):303–312.

3. Brant-Zawadzki MN, Jensen MC, Obuchowski N, Ross JS, Modic MT. Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities: a comparison of two nomenclatures. Spine (Phila Pa 1976) 1995;20(11):1257–1263.

4. 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(3):202–208.

5. Griffith JF, Wang YX, Antonio GE, et al. Modified Pfirrmann grading system for lumbar intervertebral disc degeneration. Spine (Phila Pa 1976) 2007;32(24):E708–E712.

6. 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(7):748–754.

7. Jones A, Clarke A, Freeman BJ, Lam KS, Grevitt MP. The Modic classification: inter- and intraobserver error in clinical practice. Spine (Phila Pa 1976) 2005;30(16):1867–1869.

8. Milette PC, Fontaine S, Lepanto L, Cardinal E, Breton G. Differentiating lumbar disc protrusions, disc bulges, and discs with normal contour but abnormal signal intensity: magnetic resonance imaging with discographic correlations. Spine (Phila Pa 1976) 1999;24(1):44–53.

9. Mulconrey DS, Knight RQ, Bramble JD, Paknikar S, Harty PA. Interobserver reliability in the interpretation of diagnostic lumbar MRI and nuclear imaging. Spine J 2006;6(2):177–184.

10. Peterson CK, Gatterman B, Carter JC, Humphreys BK, Weibel A. Inter- and intraexaminer reliability in identifying and classifying degenerative marrow (Modic) changes on lumbar spine magnetic resonance scans. J Manipulative Physiol Ther 2007;30(2):85–90.

11. Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine (Phila Pa 1976) 2001; 26(17):1873–1878.

12. Raininko R, Manninen H, Battié MC, Gibbons LE, Gill K, Fisher LD. Observer variability in the assessment of disc degeneration on magnetic resonance images of the lumbar and thoracic spine. Spine (Phila Pa 1976) 1995;20(9):1029–1035.

13. Smith BM, Hurwitz EL, Solsberg D, et al. Interobserver reliability of detecting lumbar intervertebral disc high-intensity zone on magnetic resonance imaging and association of high-intensity zone with pain and anular disruption. Spine (Phila Pa 1976) 1998;23(19):2074–2080.

14. Solgaard Sorensen J, Kjaer P, Jensen ST, Andersen P. Low-field magnetic resonance imaging of the lumbar spine: reliability of qualitative evaluation of disc and muscle parameters. Acta Radiol 2006;47(9):947–953.

15. 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 (Phila Pa 1976) 2002;27(10):1082–1086.

16. 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(1):299–303.

17. van Tulder MW, Assendelft WJ, Koes BW, Bouter LM. Spinal radiographic findings and nonspecific low back pain: a systematic review of observational studies. Spine (Phila Pa 1976) 1997;22(4):427–434.

18. Carrino JA, Lurie JD, Tosteson AN, et al. Lumbar spine: reliability of MR imaging findings. Radiology 2009;250(1):161–170.

19. Kovacs FM, Royuela A, Jensen TS, et al. Agreement in the interpretation of magnetic resonance images of the lumbar spine. Acta Radiol 2009;50(5):497–506.

20. Lurie JD, Doman DM, Spratt KF, Tosteson AN, Weinstein JN. Magnetic resonance imaging interpretation in patients with symptomatic lumbar spine disc herniations: comparison of clinician and radiologist readings. Spine (Phila Pa 1976) 2009;34(7):701–705.

21. Lurie JD, Tosteson AN, Tosteson TD, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine (Phila Pa 1976) 2008;33(14):1605–1610. [Published correction appears in Spine 2008;33(22):2482.]

22. Jarvik JG, Deyo RA. Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology 2009;250(1): 15–17.

23. Benoist M. The Michel Benoist and Robert Mulholland yearly European Spine Journal Review: a survey of the "medical" articles in the European Spine Journal, 2008. Eur Spine J 2009;18(1):1–12.

24. Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 1988;166(1 pt 1):193–199.

25. Arnoldi CC, Brodsky AE, Cauchoix J, et al. Lumbar spinal stenosis and nerve root entrapment syndromes: definition and classification. Clin Orthop Relat Res 1976;(115):4–5.

26. Weishaupt D, Zanetti M, Hodler J, Boos N. 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(3):661–666.

27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.

28. Lipsitz SR, Parzen M, Fitzmaurize GM, Klar N. A two-stage logistic regression model for analyzing inter-rater agreement. Psychometrika 2003;68(2):289–298.

29. Hardin JW, Hilbe JM. Generalized estimating equations. Boca Raton, Fla: Chapman & Hall, 2003.

30. Brorson S, Hróbjartsson A. Training improves agreement among doctors using the Neer system for proximal humeral fractures in a systematic review. J Clin Epidemiol 2008;61(1):7–16.

31. Modic MT, Ross JS. Lumbar degenerative disk disease. Radiology 2007;245(1): 43–61.

32. Feinstein AR, Cicchetti DV. High agreement but low kappa. I. The problems of two paradoxes. J Clin Epidemiol 1990;43(6): 543–549.