Observer Variation in MRI Evaluation of Patients Suspected of Lumbar Disk Herniation

Jeroen C. van Rijn[1]
Nina Klemetsö[2]
Johannes B. Reitsma[1]
Charles B. L. M. Majoie[2]
Frans J. Hulsmans[2]
Wilco C. Peul[3]
Jan Stam[4]
Patrick M. Bossuyt[1]
Gerard J. den Heeten[2]

OBJECTIVE. Our objective was to assess observer variation in MRI evaluation in patients suspected of lumbar disk herniation.

SUBJECTS AND METHODS. Two experienced neuroradiologists independently evaluated 59 consecutive patients with lumbosacral radicular pain. Per patient, three levels (L3–L4 through L5–S1) and the accompanying roots were evaluated on both sides. For each segment, the presence of a bulging disk or a herniation and compression of the root was reported. Images were interpreted twice: once before and once after disclosure of clinical information. Interobserver agreement was expressed as unweighted kappa values.

RESULTS. Without clinical information, interobserver agreement for the presence of herniation or bulging disk was moderate (full agreement, 84%; $\kappa = 0.63$; 95% confidence interval [CI], 0.53–0.72). Of a total of 352 segments evaluated, there was disagreement on 58 segments (17%): bulging disk versus no defect in 26 (7.4%), bulging disk versus herniation in five (1.4%), and hernia versus no defect in 27 (7.7%). With clinical information, twice as many bulging disks were reported but no new herniations were detected. Agreement slightly decreased, but not significantly (full agreement, 77%; $\kappa = 0.59$; 95% CI, 0.49–0.69; $p = 0.12$).

CONCLUSION. On average, more than 50% of interobserver variation in MRI evaluation of patients with lumbosacral radicular pain is caused by disagreement on bulging disks. Knowledge of clinical information does not influence the detection of herniations but lowers the threshold for reporting bulging disks.

The main reason for MRI referral of patients with chronic radicular pain below the knee—without a history of neoplasm, infections, or other rare abnormalities—is to distinguish between patients with and without herniated disks. This distinction requires accurate imaging because small herniations can be difficult to detect. The accuracy of MRI for predicting the presence of disk herniations at surgery is relatively high (varying from 76% to 96%) [1], and thus it has become the investigation of choice for patients suspected of lumbar disk herniations [2–4].

However, observer agreement was moderate in the few MRI studies that have reported on observer variation [5, 6]. In these, lack of clarity on the definition of bulging disk was considered a major source of observer variation.

Another potential source of variation is the influence of clinical information on image assessment. Recent textbooks on diagnostic research recommend blinding observers to reduce variation [7, 8]. In practice, most radiologists consider clinical information useful, especially in patients suspected of lumbar disk herniation. Little evidence is available on the impact of clinical information when evaluating MR images of the lumbar region.

In this prospective cohort study, we investigated observer variation in MRI evaluation for patients suspected of lumbar herniations by quantifying the amount of agreement between observers. We also assessed possible causes of disagreement between observers and the effect of clinical information on observer variation.

Subjects and Methods

Patients

We performed a prospective single-center study between June 1999 and June 2000 as part of a larger project on the diagnostic process for patients with lumbosacral radicular pain at the Amsterdam Academic Medical Centre.

Received March 8, 2004; accepted after revision June 14, 2004.

[1] Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Meibergdreef 9, Amsterdam 1105AZ, The Netherlands. Address correspondence to J. C. van Rijn (j.c.vanrijn@amc.uva.nl).

[2] Department of Radiology, Academic Medical Center, University of Amsterdam, Amsterdam 1105AZ, The Netherlands.

[3] Department of Neurosurgery, Leiden University Medical Center, Leiden 2333ZA, The Netherlands.

[4] Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam 1105AZ, The Netherlands.

AJR 2005;184:299–303

0361–803X/05/1841–299

Patients were recruited from the neurology outpatient department. Eligible were patients referred by their general practitioner with lumbosacral radicular syndrome (LRS) with suspected disk herniation at levels L3–L4, L4–L5, or L5–S1, in whom conservative treatment was unsuccessful. LRS was defined according to the national general practitioner's guideline and the consensus statement on diagnostics and treatment of LRS defined by the Dutch Neurologic Society [9]. The definition is based on continuous mono- or multiradicular pain below the knee with a primary suspicion of disk herniation. After unsuccessful conservative treatment for at least 4 weeks, patients are considered potential candidates for surgery.

Excluded were patients younger than 18 years or older than 70 years, pregnant women, patients with a previous history of lumbosacral herniation or lumbosacral surgery, and patients with contraindications for MRI.

After the neurologist had confirmed the diagnosis of LRS, patients were subjected to MRI within 1 week and no treatment was given within this period. The institutional review board approved the study protocol, and written informed consent was obtained from all patients.

Imaging Technique

Lumbar MRI examinations were performed with a 1.5-T Signa LX Scanner (GE Healthcare) using a dedicated lumbar spine surface coil. The protocol included sagittal spin-echo T1-weighted (TR/TE, 500/14) and proton-density, T2-weighted (TR 3,500/TE 120 = 20) fast spin-echo images with 4-mm slice thickness, 0.5-mm intersection gap, 200 × 512 matrix, and 29 × 29 cm field of view. In addition, axial spin-echo T1-weighted (520/12) and fast spin-echo T2-weighted (4,500/120) images were obtained from the level of L3 to the bottom of S1 with 4-mm slice thickness, 0.5-mm intersection gap, 200 × 256 matrix, and 15 × 15 cm field of view. Axial images were obtained without angulation. Finally, heavily T2-weighted (5,000/252) spin-echo oblique MR myelography was performed with two image slices of 20-mm thickness, a 250 × 220 matrix, and a 16 × 16 cm field of view.

Data Collection

Two experienced neuroradiologists evaluated all images twice within one session: once with and once without clinical information. The images were presented per patient in a random order. Directly after evaluation without clinical information, clinical information was provided in a standardized way and consisted of side, level, and severity of symptoms using a Visual Analogue Scale [10]. No other results were disclosed.

Three lumbar disks were examined per patient at levels L3–L4, L4–L5, and L5–S1. Each disk was scored for the presence of a herniation on the left or the right side. If no herniation was detected, if applicable, observers recorded the presence of a bulging disk or concluded that there was no lesion at all.

The definition of a bulging disk was according to the description by Jensen et al. [11]: "circumferential symmetric extension of the disk beyond the interspace." No distinction between protrusion and extrusion was made: Both were considered a herniated disk.

Roots corresponding with these disk levels (L4, L5, and S1) were examined per side. A five-point scale was used: definitely no root compression, possibly no root compression, indeterminate, possibly root compression, and definitely root compression (MRI examples are shown in Figs. 1A, 1B, and 2). These data were dichotomized into root compression (last two categories) and no root compression (other categories).

Statistical Analysis

Three lumbar levels were evaluated on both sides, resulting in data of six segments per patient, with two opposite segments representing one lumbar level: L3–L4 left, L3–L4 right, L4–L5 left, L4–L5 right, L5–S1 left, and L5–S1 right. For each segment, we recorded whether there was a herniated disk, a bulging disk, or no abnormality. If an observer detected a bulging disk at a certain lumbar level according to the definition described earlier, it bulged into both opposite segments. In our analysis, we considered both these segments as bulging disk although only one disk was involved. Analogously, if an observer detected a median herniation (a disk herniating into two opposite segments of one lumbar level), we also recorded both segments as a herniated disk.

First, we constructed two segment-based 3 × 3 tables (observer 1 vs observer 2) for each side of the patient (left and right). Second, we combined these two tables into one 3 × 3 table, adding the numbers in each cell. This aggregate table represents the three categories of radiologic findings (herniated disk, bulging disk, and no herniation) taking the side of the patient into account. Analogously, we constructed an aggregate 2 × 2 table for root compression.

To express observer variation, we calculated the percentage absolute agreement and the unweighted kappa values [12] using the tables described.

Bootstrap techniques [13] were used to estimate 95% confidence intervals (CIs) and to test for differences in kappa values (z test), acknowledging any correlation that might exist between segments within a single patient. Two-sided p values of less than 0.05 were used to indicate statistical significance.

All calculations were performed with SPSS 11.0 (Statistical Package for the Social Sciences) for Windows (Microsoft) and SAS 8.2 (SAS Institute) software.

Results

Sixty-four consecutive eligible patients were identified. Three patients did not undergo MRI because of claustrophobia and were excluded. Two more patients were excluded because their data were not available. Data for one patient were incomplete: Four segments were evaluated instead of six. These segments were included in the analysis.

A total of 352 segments from 59 patients (37 men and 22 women; age range, 20–70 years; mean, 44 years) were included in our analysis. Of these patients, 20 received surgical treatment for a disk herniation and three for spinal stenosis. The remaining 36 received conservative treatment.

Herniated Disks

Without clinical information, interobserver agreement on the presence of disk herniation or a bulging disk was moderate ($\kappa = 0.63$; 95% CI, 0.53–0.72; Table 1). The observers fully agreed on 294 (83.5%) of the 352 segments. In the remaining 58 segments, the observers disagreed on the presence of a herniated disk 32 times (55%). For 27 of these 32 instances, one of the observers had reported no lesion at all. For the remaining five, one observer had reported bulging disk.

After disclosure of clinical information, full agreement was reached for 272 (77%) of 352 segments. Compared with blinded evaluation, agreement was slightly lower ($\kappa = 0.59$; 95% CI, 0.49–0.69) but not significant ($p = 0.12$; Table 1).

There was a twofold increase in the total number of bulging disks reported: from 71 without clinical information to 142 with clinical information (observer 1, from 31 to 62; observer 2, from 40 to 80). Disagreement on bulging disk versus no abnormality increased from 44.8% (26/58) to 60.0% (48/80).

Disagreement on herniated disk was observed in the same 32 segments as in the blinded evaluation. However, disagreement on herniated disk versus bulging disk was observed 14 times instead of five times: After disclosure of clinical information, the diagnosis changed from no abnormality to bulging disk nine times whereas the other observer reported a herniated disk in both observations.

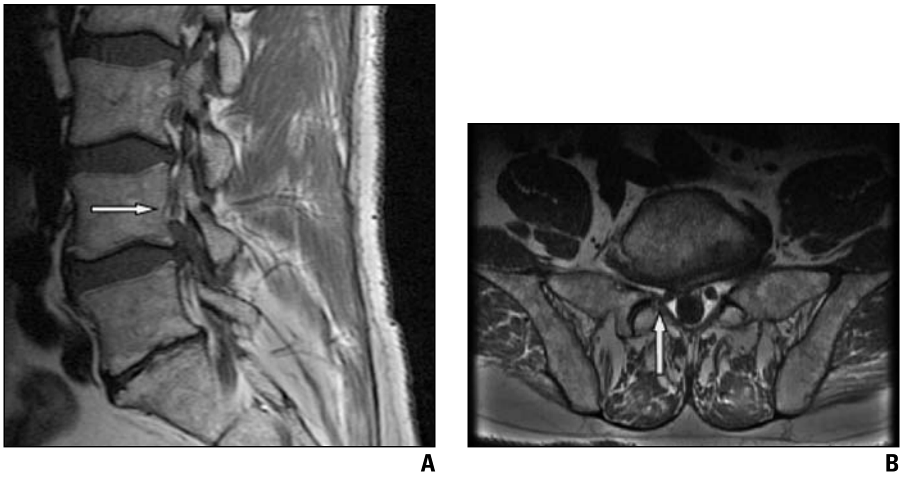

Fig. 1.—60-year-old man with right-sided lumbosacral radicular syndrome at level L5–S1.
A, Sagittal spin-echo unenhanced T1-weighted MR image (TR/TE, 500/14) shows foraminal herniation of L4–L5 disk. Arrow indicates left-sided nerve root L4, reported as "possibly no root compression" by one interpreter.
B, Axial spin-echo T1-weighted MR image (520/12) shows right-sided nerve root S1 (*arrow*), reported as "indeterminate" by both interpreters.

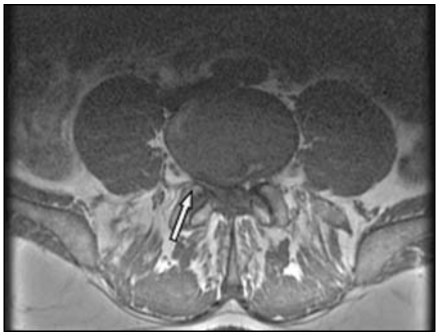

Fig. 2.—42-year-old man with left-sided lumbosacral radicular syndrome at level L4–L5. Axial spin-echo unenhanced T1-weighted MR image (TR/TE, 520/12) shows right-sided paramedian herniation of L4–L5 disk. Arrow indicates right-sided nerve root L5, reported as "possibly root compression" by both interpreters.

Median Herniated Disks

Observer 1 reported eight median herniated disks, and observer 2 reported four. With and without clinical information, both observers reported the same median herniated disks. The observers agreed on the presence of a median herniated disk four times (eight segments). Three times, observer 1 reported a median herniated disk (two segments), whereas observer 2 reported a lateral herniated disk (one segment). Once, observer 2 did not agree with observer 1 on the presence of a median herniated disk, resulting in disagreement on two segments.

Root Compression

The observers agreed fully on 329 (94%) of 352 segments without clinical information and on 326 segments (93%) with clinical information (Table 2).

Blinded and nonblinded evaluations did not differ: Without clinical information, the kappa value was 0.75 (95% CI, 0.67–0.83); with clinical information, the kappa value was 0.77 (95% CI, 0.63–0.83; $p = 0.59$).

Discussion

Our study showed substantial disagreement between observers when using MRI in patients suspected of lumbar herniations, despite its status as the gold standard. On average, we observed disagreeing results in 58 (16%) of 352 segments in 30 (51%) of 59 patients. Assessing MR images with clinical information increased the number of reported bulging disks twofold. Overall, 50% of the discordant results between the observers in our study were caused by disagreement involving the bulging-disk diagnosis, either versus no lesion or versus herniated disk.

The moderate interobserver agreement we found is well in-line with previously reported studies. In 1995, Brant-Zawadzki et al. [5] conducted an MRI study comparing two nomenclatures for lumbar herniations and concluded that bulging disk was the main reason for moderate agreement, as our findings confirmed.

In most of our patients, observers disagreed on whether a bulging disk was present or whether no abnormality was seen at all.

Bulging disks usually are assumed to be asymptomatic lesions because they are common in the general asymptomatic population (52% of asymptomatic people have at least one [11]); one could therefore argue about their clinical relevance. A second argument against the clinical relevance of bulging disks is that surgical treatment is not an option; in other words, a patient with a bulging disk and radicular pain is likely to undergo conservative treatment.

Surprisingly, we observed a twofold increase in the reporting of bulging disks after disclosure of clinical information. It seems that some small lesions are considered clinically important only when symptoms are present, although a herniated disk is not evident. Probably, radiologists lower their threshold for reporting subtle abnormalities by using the bulging-disk diagnosis as an escape option. However, the exact explanation for this influence of clinical information remains unclear.

Our study showed that clinical information does not influence the detection of herniations because both observers reported the same herniations with or without clinical information. Still, disagreement on the presence of a herniated disk remains a key issue. On average, we observed disagreement on the presence of a herniated disk for 10% of all segments in 21 (36%) of 59 patients. This type of disagreement is clinically relevant, because the decision whether to surgically intervene depends greatly on a clear MRI diagnosis. In practice, uncertainty on the presence of a herniated disk will—in most cases—result in conservative treatment, just or unjust.

To obtain better insight into the possible causes of discrepancies, a panel reevaluated the MR images involving the presence of a herniated disk. For 32 segments of 21 patients, the observers disagreed on the presence of a herniated disk (Table 1). Additional pathology was considered to cause disagreement in the evaluation of seven segments (five patients): spondylolisthesis (three segments; two patients), collapsed vertebrae (two segments; one patient), or annular tear (two segments; two patients).

For five segments, no additional pathology was reported, but a median herniation caused disagreement: Because of the inability to choose between left and right, a median herniated disk was scored positive in two opposite segments (the equivalent of one lumbar level). For three segments, one of the observers reported a median herniation (two positive segments), whereas the other observer reported either a left- or a

TABLE 1 Disk Evaluation

Radiologist 2	Radiologist 1							
	Blinded Evaluation				Evaluation with Clinical Information			
	HNP	Bulge	No.	Total	HNP	Bulge	No.	Total
HNP	50	2	17	69	51	8	11	70
Bulge	3	20	17	40	6	40	34	80
No.	10	9	224	243	7	14	181	202
Total	63	31	258	352	64	62	226	352
Agreement	Full = 84% (294/352) κ = 0.63 (95% CI, 0.53–0.72)				Full = 77% (272/352) κ = 0.59 (95% CI, 0.49–0.69; $p = 0.12$)[a]			
Disagreement								
No. vs bulge (%)	26 (44.8)				48 (60.0)			
Bulge vs HNP (%)	5 (8.6)				14 (17.5)			
HNP vs no. (%)	27 (46.6)				18 (22.5)			
Total (%)	58 (100)				80 (100)			

Note.—HNP = Hernia nuclei pulposi, CI = confidence interval.

[a] Difference in kappa of blinded evaluation vs kappa of evaluation with clinical information.

TABLE 2 Root Compression

Radiologist 2	Radiologist 1					
	Blinded Evaluation			Evaluation with Clinical Information		
	Yes	No	Total	Yes	No	Total
Yes	44	11	55	45	9	54
No	12	285	297	17	281	298
Total	56	296	352	62	290	352
Agreement	Full = 94% (329/352) κ = 0.77 (95% CI, 0.63–0.83)			Full = 93% (326/352) κ = 0.75 (95% CI, 0.67–0.83; $p = 0.59$)[a]		

Note.—CI= confidence interval.

[a] Difference in kappa of blinded evaluation vs kappa of evaluation with clinical information.

right-sided herniation. In other words, both observers detected the same herniated disk at the same level, but they disagreed on its location: median or lateral. For the other two segments, disagreement occurred because one of the observers reported a median herniated disk, whereas the other observer recorded no lesion.

For three segments (two patients), "no clarity on the level of L5–S1" was considered to cause disagreement, and for 17 segments, the small size of the herniation was considered to be the cause of disagreement.

Our study showed clinical information to have virtually no influence on the assessment of root compression. Although 93% full agreement seems high, this figure was highly influenced by the low prevalence of root compression. The accompanying kappa values are disappointingly moderate, considering MRI to be the reference standard. However, our results are in-line with previously published data [14, 15] although these studies did not use a dichotomized five-point scale.

Another point of discussion is that, without a clear reference standard, it remains unclear whether MRI can always depict all cases of root compression.

The purpose of our study was to consecutively include all patients in whom herniated disk was clearly suspected but secondary neoplasms, infections, or other rare causes were not. However, a potential bias was introduced in our study during the patient selection procedure. After referral by their general practitioner, outpatients were selected, but the exact number of patients who were referred but not selected was not recorded. We assume this situation applies to very few patients.

A general problem in imaging studies on lumbar herniations and root compression is the lack of a proper reference standard. Surgery is not satisfactory because surgical exploration of all segments is unethical. Follow-up is also not satisfactory because a major part of these herniations is self-limiting. That is why we decided to refrain from an accuracy study. The scope of this study was observer variation in radiologic findings only. We also refrained from trying to retrieve the cause of symptoms or to distinguish between asymptomatic and symptomatic lesions by correlating clinical findings with MRI. A reliable reference standard is needed to conduct that kind of study.

Another limitation was the relatively few observers and patients involved, restricting the statistical power to detect a significant difference between the blinded and nonblinded evaluations. However, because most studies on this subject have had limited numbers of observers and patients, our results can be compared with those results.

Observer variation is a general problem in research settings. Clear methodology on how to conduct a proper observer-variation study in imaging research is not available and should be developed. For now, in our opinion, diagnostic-imaging studies should focus not just on reporting agreement measures but also on investigating possible causes of disagreement.

In summary, although MRI is the investigation of choice for the evaluation of patients suspected of having herniated disks, substantial observer disagreement was observed. Disagreement was caused mainly by a lack of consensus on bulging-disk diagnosis. After disclosure of clinical information, disagreement was even higher—twice as many bulging disks were reported. Detection of herniated disks did not differ between MRI evaluations performed with and without clinical information.

As long as no definite consensus exists on the nomenclature and clinical relevance of bulging disks, we advise radiologists to adopt a high threshold for reporting them. We found no benefit from having elaborate clinical information at hand for the MRI evaluation of herniated disks.

References

1. Jarvik JJ, Hollingworth W, Heagerty P, et al. The Longitudinal Assessment of Imaging and Disability of the Back (LAIDBack) study. *Spine* 2001;26:1158–1165
2. Patel N. Surgical disorders of the thoracic and lumbar spine: a guide for neurologists. *J Neurol Neurosurg Psychiatr* 2002;73[suppl I]:42–i48
3. Milette PC. Classification, diagnostic imaging, and imaging characterization of a lumbar herniated disk. *Radiol Clin North Am* 2000;38:1267–1292
4. Herzog RJ. The radiologic assessment for a lumbar disc herniation. *Spine* 1996;21[suppl]:19S–38S
5. 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* 1995;1:20:1257–1263
6. Raininko R, Manninen H, Battie MC, Gibbons ME, Gill K, Fisher LD. Observer variability in the assessment of disc degeneration on magnetic resonance images of the lumbar and thoracic spine. *Spine* 1995;120:1029–1035
7. Pepe MS. *The statistical evaluation of medical tests for classification and prediction.* Oxford, UK: Oxford University Press, 2003:5–7
8. Zhou X, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine.* New York, NY: Wiley, 2002:86–88
9. Stam J. Consensus on diagnosis and treatment of lumbosacral root entrapment syndromes [in Dutch]. *Ned Tijdschr Geneeskd* 1996;140:2621–2627
10. Streiner DL, Norman GR. *Health measurement scales,* 3rd ed. Oxford, UK: Oxford University Press, 2003:32–34
11. Jensen MC, Brant-Zawadski MN, Obuchowski N, et al. Magnetic resonance imaging of the lumbar spine in people without back pain. *N Engl J Med* 1994;331:69–73
12. Fleiss JL. *Statistical methods for rates and proportions,* 2nd ed. Hoboken, NJ: John Wiley and Sons, 1981:225–232
13. Efron B, Tibshirani RJ. *An introduction to the bootstrap.* Boca Raton, FL: Chapman and Hall, 1993
14. Vroomen PC, De Krom MC, Wilmink JT. Pathoanatomy of clinical findings in patients with sciatica: a magnetic resonance imaging study. *J Neurosurg* 2000;92[2 suppl]:135–141
15. Vroomen PC, De Krom MC. Knottnerus JA. When does the patient with a disc herniation undergo lumbosacral discectomy? *J Neurol Neurosurg Psychiatr* 2000;68:75–79