Value of the bell test and the hyperextension test for diagnosis in sciatica associated with disc herniation: comparison with Lasègue’s sign and the crossed Lasègue’s sign

S. Poiradeau, V. Foltz, J.-L. Drapé1, J. Fermanian2, M. M. Lefèvre-Colau, M. A. Mayoux-Benhamou and M. Revel

Service de rééducation et de réadaptation de l’appareil locomoteur et des pathologies du rachis, 1Service de radiologie ostéo-articulaire, Assistance Publique-Hôpitaux de Paris, Université René Descartes, Paris and 2Hôpital Necker-Enfants Malades, Service de Biostatistiques, Assistance Publique-Hôpitaux de Paris, Université René Descartes, 149–161 Rue de Sèvres, 75015 Paris, France

Abstract

Objectives. To evaluate the reliability, sensitivity, specificity and positive (PPV) and negative (NPV) predictive values for the diagnosis of sciatica associated with disc herniation of the bell test (BT) and the hyperextension test (HT).

Methods. According to magnetic resonance imaging, computed tomography scanning or myelography findings, patients were classified as having sciatica associated with disc herniation (group A) or sciatica without disc herniation or sciatica of other mechanical origin (group B). Four clinical manoeuvres [bell test (BT), hyperextension test (HT), Lasègue’s sign (LS) and the crossed Lasègue’s sign (CL)] were tested by three investigators. Intra- and interobserver reliabilities were calculated using the κ correlation coefficient or the intraclass correlation coefficient (ICC). The sensitivity, specificity, PPV and NPV of the four manoeuvres were calculated. Stepwise logistic regression analysis was performed to determine the best set of variables predicting sciatica caused by disc herniation.

Results. Seventy-eight patients (43 in group A, 35 in group B; 33 males) with a mean age of 50 ± 16 yr were included. Interobserver reliabilities ranged from 0.58 to 0.64 for the BT, 0.35 to 0.50 for the HT, 0.27 to 0.47 for LS and 0.43 to 0.72 for CL. LS had the best sensitivity (0.77–0.83) and CL the best specificity (0.74–0.89), while PPV and NPV were equivalent for the four manoeuvres (0.55–0.75 for PPV and 0.45–0.59 for NPV). The best PPV was observed for the association of HT with CL (0.67–0.85). Stepwise logistic regression analysis did not allow us to propose a set of variables predicting sciatica caused by disc herniation.

Conclusion. This study suggests that clinical values of the BT and HT are of interest, and are similar to those of LS and CL.

Key words: Sciatica, Disc herniation, Clinical examination, Reliability, Positive predictive value, Negative predictive value, Sensitivity, Specificity.

Sciatica is among the most common musculoskeletal disorders, with an estimated prevalence of around 5% [1]. Sciatica is usually defined as pain from the back radiating to the buttocks and down the leg as far as the knee, ankle or foot [1]. Radicular pain usually results from various mechanical lesions, such as disc herniation, spinal stenosis, facet joint degeneration or synovial cysts, and congenital anomalies. Various mechanisms have been implicated in radicular pain, but patients are usually classified as having sciatica associated with nerve root compression (also called true sciatica) or sciatica without nerve root compression (also called referred pain).

Modern imaging methods using magnetic resonance imaging (MRI) and computed tomography (CT) have
improved the diagnosis of lumbar spinal syndromes but have not abolished the need for thorough clinical evaluation. History and physical examination, in tandem with radiological findings, provide the basis for the diagnosis of sciatica caused by a herniated lumbar disc. However, few clinical signs used in sciatica have been evaluated for their diagnostic value and their reliability. Published studies mostly concern Lasègue’s sign (LS) and the crossed Lasègue’s sign (CL) [2–14]. Testing for Lasègue’s sign is performed by raising passively the leg affected with sciatica, with the knee in full extension. The sign is positive if sciatica is reproduced or exacerbated. The test for CL is similar to that for LS, but the contralateral leg is raised passively and the sign is positive if sciatica is reproduced or exacerbated. These two signs seem to be reliable and to have good specificity (CL) or sensitivity (LS) in the diagnosis of disc herniation [3–5, 7–9, 11–14]. However, most trials have been performed retrospectively [4, 12] on patients scheduled for discectomy [2, 4, 5, 7, 8, 11, 12] by examiners who were aware of the diagnosis [2, 7, 11, 13]. Moreover, the populations studied were clinically heterogeneous, with patients suffering from sciatica (L5 or S1), other radicular pain (L2, L3, L4) or isolated low back pain [4, 5, 7, 9, 11, 12, 14]. In addition, some authors consider the test result normal when radicular pain occurs above 70° [11, 15, 16], whereas others define a positive result as the occurrence or exacerbation of any lumbar or radicular pain during straight leg raising [12, 13].

Several other manoeuvres are used widely in the clinical examination of patients with sciatica. Among these, the reproduction of radicular pain by pressure applied with the thumb between spinous processes L4 and L5 or between L5 and S1 and in the corresponding paravertebral area (called the bell test by analogy with ringing a bell), described by De Séze in 1946 [17], and by lumbar hyperextension (HE) are two clinical signs for which sensitivity, specificity and positive (PPV) and negative (NPV) predictive values for the diagnosis of sciatica associated with disc herniation have never been studied. The original study on the bell test (BT) concluded that the bell sign had a diagnostic and localization value for disc herniation [17], but the reliability of the BT was not reported and the results of this study was never confirmed. The significance of pain exacerbation during lumbar hyperextension has been poorly studied and is not fully understood. Finally, the question of the role of the examiner’s experience in the diagnostic value of these clinical signs has never been raised.

Thus, evidence that a particular clinical presentation can discriminate sciatica associated with disc herniation from sciatica of other mechanical origin is lacking. The aim of this study was to compare the reliabilities and diagnostic values of two clinical manoeuvres widely used and evaluated (LS, CL) with those of two less well known tests [BT, hyperextension test (HT)]. We assessed: (i) test–retest and interobserver reliability for BT, HE, LS and CL; (ii) the sensitivity, specificity, PPV and NPV of these four clinical signs, considered separately or in combination, for the diagnosis of sciatica associated with disc herniation.

Patients and methods

Patient selection

Patients hospitalized for acute or chronic sciatica of mechanical origin were included consecutively in this prospective study. Sciatica was defined as having lumbosacral and lower limb pain, associated or not with paraesthesias and with one of the following conditions: radicular pain below the knee after an L5 or S1 nerve root dermatome; and radicular pain above the knee associated with neurological impairment (reflex abolition, muscular weakness or sensory defects in the corresponding radicular area).

Exclusion criteria were: (i) low back pain without sciatica; (ii) radicular pain in a dermatome other than L5 and/or S1; (iii) systemic lumbosacral pain (tumour, infectious or inflammatory disease); (iv) prior lumbar surgery; and (v) uncontrolled psychiatric disorder.

Investigators

Three investigators participated in this study. The first (E1) was a rheumatology trainee with 3 yr of experience, who worked occasionally in the department. The second (E2) and third examiners (E3) were full-time physicians who had worked in the department for 10 and 25 yr respectively.

Physical examination

Bell test. This was performed with the patient in the standing position. The test was positive when the examiner reproduced or exacerbated the usual radicular pain by pressure applied with the thumb between the spinous processes L4 and L5 or between L5 and S1, or in the near corresponding paraspinal area. When the manoeuvre reproduced only lumbar pain, it was considered negative.

Hyperextension test. This was performed with the patient standing. The trunk was mobilized passively and slowly over the full range of extension with the knees in extension. The test was positive if the sciatica was reproduced or worsened. If the manoeuvre was interrupted because of lumbar pain, it was considered negative.

Lasègue’s sign. This was investigated with the patient supine. The leg affected with sciatica was slowly raised passively, with the patient relaxed and the knee in full extension. Elevation was stopped when the patient began to feel pain. The sign was positive only if sciatica was reproduced or exacerbated. If the manoeuvre was interrupted because of lumbar pain or hamstring stiffness, it was considered negative. When the test was positive, the angle of elevation was recorded using a goniometer. No limiting angle was defined.
Crossed Lasègue’s sign. This was performed in the same conditions as the LS but the contralateral leg was passively raised. The sign was positive only if sciatica was reproduced or exacerbated. No limiting angle was defined.

Clinical procedure
E1 first checked that patients fulfilled the inclusion criteria. The patient was then informed about the procedure and gave written agreement to participation. Demographic and clinical characteristics were recorded, including age, sex, work injury, pain duration, radicular pain topography, low back and radicular pain intensities on a 100 mm visual analogue scale (VAS), and the presence of neurological impairment (muscular strength, sensory defect, reflex disturbance, cauda equina syndrome).

Within the same day, the three observers examined each patient blindly and independently of each other. At the time of examination, each examiner was unaware of the diagnosis and of the radiological findings. E1 was always the first examiner. The sequence of examination by E2 and E3 was chosen at random. Each patient had five examinations performed successively within 12 h. E1 and E2 each performed two examinations for the test–retest reliability study. Each observer collected all data in separate booklets.

Radiographic assessment
E1, E2 and an experienced radiologist analysed blindly and independently at least one of the three following lumbar spine radiological investigations, as available: MRI, CT scanning and saccoradiculography. These investigations were not prescribed for the study but for diagnosis and treatment decisions. According to the decision of the three readers, the patients were divided into two groups: compression of the nerve root by disc herniation (group A); sciatica without disc herniation or associated with degenerative changes or spinal stenosis (group B) [18, 19].

When the three readers disagreed on patient classification, radiological images were analysed blindly by E3. If a majority (3 to 1) was not obtained, the patient was not included in the study.

Statistical analysis
Data were analysed using Systat (Chicago, IL, USA) 5.2.1 software. The qualitative data were described using percentages and proportions. The quantitative results are expressed as mean ± standard deviation (s.d.) and range. For qualitative values, the intra- and intertester reliabilities were calculated as the $\kappa$ coefficient of agreement. For E1 and E2, who each performed two examinations, the mean $\kappa$ coefficients are presented. For quantitative values, the intra- and intertester reliabilities were calculated as the intraclass coefficient of correlation (ICC). The sensitivity, specificity, PPV and NPV were calculated for each sign and for their combinations. Stepwise logistic regression was performed using the SAS software, as follows: explanatory variables were BT, HE, LS, CL; and several logistic regression models were tested using a maximum likelihood method of compilation. Validity of fitting was assessed taking into account the results of the $\chi^2$ test. The $P$ levels for entry and removal of predictors were set at 0.10 and 0.15 respectively. The overall discriminating power of the model was assessed using predicted probabilities and observed responses, and the area under the receiver operator characteristic curve.

Results
Demographic and clinical data
Data are summarized in Table 1. Over a 1 yr period, 78 patients (33 males, 45 females; mean age 50 ± 16 years, range 17–85 yr) were included. Forty three patients were included in group A (21 males, mean age 44 ± 12 yr, range 19–73 yr) and 35 in group B (12 males, mean age 58 ± 18 yr, range 17–85 yr).

Radiographic analysis
Agreement among the three readers was obtained for 71 patients (92%). After analysis by the fourth reader (E3), two patients were included in group A and four in group B. In one case no agreement was reached and the patient was not included.

Reliability of the four signs
Intraobserver reliability. Intraobserver reliability was good (0.77 and 0.94 for BT, 0.76 and 0.96 for HT, 0.8 and 0.95 for LS, and 0.77 and 0.93 for CL for E1 and E2 respectively).

Interobserver reliability. Results are summarized in Table 2.

Bell test. The BT was positive in 29 patients for E1 (16 in group A, 13 in group B), 35 patients for E2 (21 in group A, 14 in group B) and 36 patients for E3 (23 in group A, 13 in group B). Inter-rater reliability was fair in each group of patients and for each pair of examiners ($\kappa = 0.58–0.64$).

Hyperextension test. This test was positive in 26 patients for E1 (17 in group A, nine in group B), 34 patients for E2 (20 in group A, 14 in group B) and 31 patients for E3 (21 in group A, 10 in group B). Inter-rater reliability was weak to fair in each group of patients and for each pair of examiners ($\kappa = 0.35–0.50$).

Lasègue’s sign. This manoeuvre was positive in 55 patients for E1 (33 in group A, 22 in group B), 58 patients for E2 (36 in group A, 22 in group B) and 58 patients for E3 (34 in group A, 24 in group B). Inter-rater reliability was weak ($\kappa = 0.27–0.47$). The E1–E2 pair had better reliability than the other pairs.

Crossed Lasègue’s sign. This manoeuvre was positive in 12 patients for E1 (nine in group A, three in group B),


21 patients for E2 (14 in group A, seven in group B), and 19 patients for E3 (14 in group A, five in group B). Inter-rater reliability was weak for pairs E1–E2 and E1–E3 ($\kappa = 0.48$ and $0.43$ respectively). The reliability for the E2–E3 pair was good ($\kappa = 0.72$).

**Interobserver reliability of the degree of Lasègue’s sign**

The degree of LS was assessed without any measurement tool. Mean values were $63 \pm 54^\circ$ (range 10–90°) for E1, $66 \pm 59^\circ$ (range 15–110°) for E2 and $67 \pm 62^\circ$ (range 20–100°) for E3. The ICC values were $0.78$ for E1–E2, $0.58$ for E1–E3 and $0.80$ for E2–E3.

**Sensitivity, specificity, and positive and negative predictive values of the four signs**

The sensitivity, specificity, PPV and NPV of BT, HE, LS and CL for the three examiners are summarized in Table 3. For examiners E1 and E2, the data presented are means of values obtained during the two examinations.

**Sensitivity.** LS had the best sensitivity for the diagnosis of sciatica associated with disc herniation ($0.77–0.83$). The sensitivities of BT, HE and CL were lower ($0.21–0.53$). For each sign, there was no difference among the three examiners.

**Specificity.** CL had the best specificity ($0.74–0.89$). BT and HE had fair to good specificity ($0.62–0.63$ and $0.59–0.72$ respectively) and that of LS was weak ($0.36–0.39$). There was no difference among the three examiners.

**PPV.** The PPVs of the four signs were fair ($0.55–0.69$) except for E3, for whom the PPV of CL was good ($0.75$).

**NPV.** The NPVs of the four signs were weak to fair ($0.45–0.63$), and were similar for the three examiners.

**Sensitivity, specificity PPV and NPV of sign combinations**

Results are summarized in Table 4. Sensitivities of sign combinations were always weaker than the sensitivity of each of the signs. The best specificity ($0.92–0.94$) and PPV (up to 0.85) were observed with the HE–CL combination. NPVs were not modified by the combination of signs.

**Stepwise logistic regression analysis**

Stepwise logistic regression analysis did not find a set of variables predicting sciatica caused by disc herniation. Only LS for E2 was at the limit of significance ($\chi^2 = 0.04$). With this variable, 57% of patients were classified correctly. The area under the curve was 0.64.

**Discussion**

The gold standard for the diagnosis of disc herniation in this work was the radiological findings rather than surgical verification. Myelography, CT scanning and MRI have been shown to be accurate for the detection of disc herniation, being in agreement with surgical verification in 81% of cases for myelography, 93% for CT scanning and between 91 and 96% for MRI [18, 20, 21]. Moreover, the choice of radiological findings as the gold standard has the advantage of allowing the inclusion of patients with differing pain intensities, pain durations and disability rather than only those requiring surgery.

We show that the interobserver reliabilities of four clinical signs used in daily practice are at best fair. This lack of reliability may reflect not only interobserver variation but also daily spontaneous fluctuations of symptoms or the fluctuating effects of analgesic drugs,

<table>
<thead>
<tr>
<th><strong>Table 1. Demographic and clinical characteristics of the patients</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All patients</strong></td>
</tr>
<tr>
<td>Number of patients</td>
</tr>
<tr>
<td>Age ± S.D. (yr)</td>
</tr>
<tr>
<td>Sex M/F</td>
</tr>
<tr>
<td>Work injury (no. patients)</td>
</tr>
<tr>
<td>Pain duration ± S.D. (days)</td>
</tr>
<tr>
<td>Lumbar pain (VAS) ± S.D. (range 0–100)</td>
</tr>
<tr>
<td>Radicular pain (VAS) ± S.D. (range 0–100)</td>
</tr>
<tr>
<td>Neurological impairment (no. patients)</td>
</tr>
<tr>
<td>M = male; F = female.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Table 2. Inter-rater reliability of BT, HT, LS and CL (κ coefficient)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pair of investigators</strong></td>
</tr>
<tr>
<td>E1</td>
</tr>
<tr>
<td>E2</td>
</tr>
<tr>
<td>Group A</td>
</tr>
<tr>
<td>Group B</td>
</tr>
</tbody>
</table>
as proposed by Kosteljanetz et al. [8]. As we tried to stay as close as possible to the conditions found in clinical practice, the three observers made their clinical examinations at intervals ranging from several minutes to several hours in order to take into account variations in appointment schedules in clinical practice as well as diurnal variations in each sign [22]. This was not the case for the intraobserver reliability study, in which E1 and E2 performed their two examinations at intervals of a few minutes. The higher reliability values observed surely reflect the weaker intraobserver than interobserver variations, but may also have resulted from the absence of diurnal variability in this situation.

Interobserver reliabilities for the BT and HT are reported for the first time in this study. The values observed were of the same order as those found for CL and slightly better than those observed for LS, suggesting that the bell and hyperextension signs could be included in the systematic clinical assessment of patients with radicular pain. Moreover, the sensitivities, specificities and positive predictive values of the HT and BT are of interest. Compared with CL, their sensitivities were better and their specificities and PPV were equivalent. The BT was first described in 1946 by De Sèze [17] and was considered to be of value for the localization of disc herniation. In fact, this author reported concordance between the test and surgical findings in 90% of patients for the presence of disc herniation as well as for the localization of the herniated disc (L4–L5 disc herniation when the test was positive at the L4–L5 level and L5–S1 disc herniation when the test was positive at the L5–S1 level). The significance of pain exacerbation during lumbar hyperextension has been poorly studied in low back and radicular pain and is not fully understood [23–25]. Two retrospective studies have shown that the ability of patients with herniated nucleus pulposus and sciatica to achieve full passive lumbar extension was a useful predictor of a favourable response to conservative management [26, 27].

The values of interobserver reliabilities for LS reported here are in accordance with most published studies [9, 13, 14]. Waddel et al. [14] found an agreement coefficient of $\kappa = 0.56$. McCombe et al. [9] reported an agreement coefficient between two orthopaedic surgeons and one orthopaedic surgeon and one physiotherapist of $\kappa = 0.66$ and 0.44 respectively. In the latter study, as in ours, LS was considered positive when straight leg raising reproduced leg pain. More recently, an agreement coefficient of $\kappa = 0.93$ between trained examiners has been reported [13]. However, this study included only subjects with low back pain but without radicular pain and the positivity of the manoeuvre was not defined.

In several studies, LS is considered negative if pain is not reproduced before passive leg raising of less than 75° [14]. This restriction does not take into account very supple patients, in whom typical pain can be reproduced by straight leg raising through more than 90°. Moreover, it has been shown that this restriction is insufficient to increase the reliability of LS [28].

The interobserver reliability values of the degree of LS were good except for the pair E1–E3. Conflicting data
have been reported on this subject [8, 9, 29–31]. These studies are difficult to compare because they included different groups of subjects with different statistical analyses.

The κ coefficients reported in our study for CL are lower, except for one of the three pairs of examiners, than that observed by McCombe et al. [9] between two orthopaedic surgeons (κ = 0.74). However, in the latter trial the κ coefficient between an orthopaedic surgeon and a physiotherapist was close to zero (−0.02). Therefore, this manoeuvre, which could be considered very simple, can be interpreted or performed differently according to the examiner’s habit.

Using radiological findings as the gold standard, we found that the sensitivities and specificities of LS and CL for the diagnosis of nerve root compression by disc herniation were similar to those observed in other trials using surgical verification [2, 4, 5, 8, 11, 12, 15]. We also report fair or good PPV for LS and CL. A comparable PPV value (0.64) has been reported for LS in patients undergoing surgery for herniated disc, but a better PPV value was found for CL (0.96) [5].

Combinations of signs did not improve sensitivity or NPV for the diagnosis of sciatica associated with disc herniation. Nevertheless, the combination of HT with CL had a high specificity (0.92–0.94) and increased the PPV to 0.85. This association was observed in 25% of patients with disc herniation, explaining why logistic regression failed to select a set of variables allowing correct classification of patients. However, this result demonstrates the value of a simple clinical examination in sciatica.

Finally, we show for the first time that the reliability, sensitivity, specificity, PPV and PNV of these four clinical signs are not dependent on the examiner’s experience. However, this observation must be tempered by two factors: first, these manoeuvres are easy to perform and easily learned; secondly, the youngest examiner was a physician specializing in rheumatology. These results will have to be confirmed by studies involving medical students, general practitioners and spine specialists.

In conclusion, this study suggests that the clinical values of the BT and HE are of interest and at least similar to those of LS and CL. The combination of HT with CL has excellent specificity and a good PPV for the diagnosis of sciatica associated with disc herniation. Thus, the BT and HT could be performed systematically in standardized physical examination of sciatica.

References