Eur Spine J (2016) 25:2734–2740
DOI 10.1007/s00586-015-4359-2

ORIGINAL ARTICLE

Reliability among clinicians diagnosing low back-related leg pain

Siobhán Stynes[1] · Kika Konstantinou[1] · Kate M. Dunn[1] · Martyn Lewis[1] · Elaine M. Hay[1]

Received: 5 June 2015 / Revised: 7 December 2015 / Accepted: 7 December 2015 / Published online: 24 December 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract
Purpose To investigate agreement and reliability among clinicians when diagnosing low back-related leg pain (LBLP) in primary care consulters.
Methods Thirty-six patients were assessed by one of six physiotherapists and diagnosed as having either leg pain due to nerve root involvement (sciatica) or referred leg pain. Assessments were video recorded. In part one, the physiotherapists each viewed videos of six patients they had not assessed. In part two, videos were viewed by another six health professionals. All clinicians made an independent differential diagnosis and rated their confidence with diagnosis (range 50–100 %).
Results In part one agreement was 72 % with fair inter-rater reliability ($K = 0.35$, 95 % CI 0.07, 0.63). Results for part two were almost identical ($K = 0.34$, 95 % CI 0.02, 0.69). Agreement and reliability indices improved as diagnostic confidence increased.
Conclusion Reliability was fair among clinicians from different backgrounds when diagnosing LBLP but improved substantially with high confidence in clinical diagnosis.

Keywords Sciatica · Reliability · Differential diagnosis · Low back-related leg pain

✉ Siobhán Stynes
s.stynes@keele.ac.uk

[1] Arthritis Research UK Primary Care Centre, Research Institute for Primary Care and Health Sciences, Keele University, Keele ST5 5BG, Staffordshire, UK

Introduction

Low back-related leg pain (LBLP) can be classified as either radicular pain due to nerve root involvement (NRI) or referred (non-specific) pain due to back pain spreading down the leg (from structures such as ligament, joint or disc but not involving a spinal nerve root). The clinical task of differentiating NRI from referred leg pain in LBLP patients is recognized as important in lines with clinical guidelines [1], but can be difficult in clinical practice [2–4].

Although the diagnosis of NRI is predominantly clinical, there is no accepted diagnostic "gold standard". Items from history [5] and physical examination [6] in patients with nerve root symptoms due to disc herniation have mostly shown poor individual diagnostic performance. Many of the studies have been carried out in secondary care and have often used magnetic resonance imaging (MRI) as the reference standard [6]. However, the usefulness of MRI as a reference test has been questioned. Positive MRI findings can be found in asymptomatic people [7], patients with nerve root symptoms can have normal MRIs [8] and MRI findings fail to distinguish sciatica patients in terms of the symptom severity [9]. Literature suggests that in the absence of a well-accepted reference standard, expert clinical opinion may be considered an appropriate alternative for diagnosis, providing that it is reasonably reliable [10].

Reliability of individual clinical tests to identify NRI has been documented as mainly poor [6] and agreement on self-reported features of NRI has generally not shown better than fair reliability [11, 12]. However, the reliability of the overall decision as to whether a clinical presentation in LBLP patients is NRI or referred pain has received less attention. One study that did investigate this showed considerable inter-rater variability among neurologists when

Table 1 Inclusion and exclusion criteria

Inclusion criteria
Back-related leg pain of any duration and severity
Exclusion criteria
'Red flags' indicative of possible serious spinal pathology
Previous lumbar spinal surgery
Serious co-morbidity or mental health problems
Pregnancy
Currently receiving physiotherapy, osteopathy or chiropractic treatment
Under a secondary care doctor for the same problem
Unable to read or speak English

asked to identify the presence of NRI based on history and physical examination in patients with LBLP [13].

Despite the recognized importance of differentiating between NRI and referred spinal pain to inform clinical management [14], there is a lack of studies examining the reliability of this diagnostic decision. The aim of this study was therefore to investigate the agreement and reliability among clinicians when diagnosing patients presenting in primary care with symptoms of LBLP. Agreement is the degree to which ratings are identical. Reliability is agreement beyond chance and reflects the ratio of variability between ratings of the same subjects to the total variability of all ratings in the sample [15]. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [15] were followed in this report.

Methods

There were two parts to this study. In part one, the raters were trained experienced musculoskeletal physiotherapists. They each carried out assessments on LBLP patients which were video recorded, then at a later date they watched video assessments of patients not examined by them. These physiotherapists are named Group A when assessing the patients and Group B when watching the patients' video assessments. In part two, a group of health care professionals (who had not participated in the assessments) from varied clinical backgrounds, watched the same patients assessments on video. The aim for part two was to gain a broader insight into current agreement on the clinical diagnosis of LBLP among health care professionals. These raters are named Group C.

Sample

Subjects were recruited as a sample of convenience from participants in an observational cohort study of primary care consulters with LBLP (the ATLAS study). Details of the ATLAS study protocol are reported elsewhere [16]. In brief, patients visiting their General Practitioner (GP) with LBLP were invited to attend a research clinic where they underwent a clinical assessment by a physiotherapist. The inclusion and exclusion criteria for the main cohort (ATLAS) are detailed in Table 1 and apply to the reliability study as well. Patients who agreed to take part in the reliability study consented to allow their clinical assessment to be video recorded. Ethical approval was granted by the South Birmingham Research Ethics Committee. Recruitment to the reliability study took place from August 2011 to July 2012.

For this two rater inter-rater agreement and reliability evaluation, at least 30 subjects were needed for analysis at 90 % power to detect a statistically significant kappa of 0.6 [from a null hypothesis value of 0 ($\alpha = 0.05$)] with a 95 % confidence interval (CI) [17].

Raters and training

The raters for part one were the six physiotherapists involved in the ATLAS research clinics. As part of the ATLAS study they attended training sessions related to the procedures of the study. Details of the training are reported elsewhere [16]. Raters for part two were six health professionals involved in managing LBLP patients. They did not participate in any prior training.

Assessment

The clinical assessment for LBLP was developed following consensus from a Delphi study involving representatives from low back pain disciplines [18]. The clinical history questions and physical examination items used were the same as those described in low back pain (LBP) guidelines and specialty books.

Part one

The physiotherapists (Group A) completed the clinical assessment which took approximately 30 min and was video recorded. At the end of the assessment they answered two written questions (Box 1) relating to diagnosis and diagnostic confidence.

Box 1 Questions clinicians answered at the end of the assessment

1. Is this low back pain with nerve root involvement: Yes / No
2. How confident are you in your clinical impression (rate on a 0-100% scale where 100% means absolutely certain/confident)?

At a later date, each physiotherapist (Group B) watched videos of six patients they had not assessed. The order in which they viewed the videos was not predetermined and they answered the same questions (Box 1). They did not have access to the clinical notes made by the assessing physiotherapist and were blind to that assessor's diagnostic decision. The videos had been edited to remove any dialogue between the patient and therapist where assessment findings and diagnosis were discussed.

Part two

The health professionals involved in part two (Group C) each watched six videos and answered the same two questions (Box 1). The study flow chart is shown in Fig. 1.

Data analysis

Results were summarized using percentage agreements and kappa coefficients with two sided 95 % CIs. Kappa coefficients were computed using SPSS version 20. Interpretation of the kappa coefficient was used whereby kappa 0–0.2 indicates slight agreement; 0.21–0.4 fair agreement; 0.41–0.6 moderate agreement; 0.61–0.8 substantial agreement and 0.81–1.0 almost perfect agreement [19]. The relationship between different levels of diagnostic confidence and agreement and reliability indices was also reported.

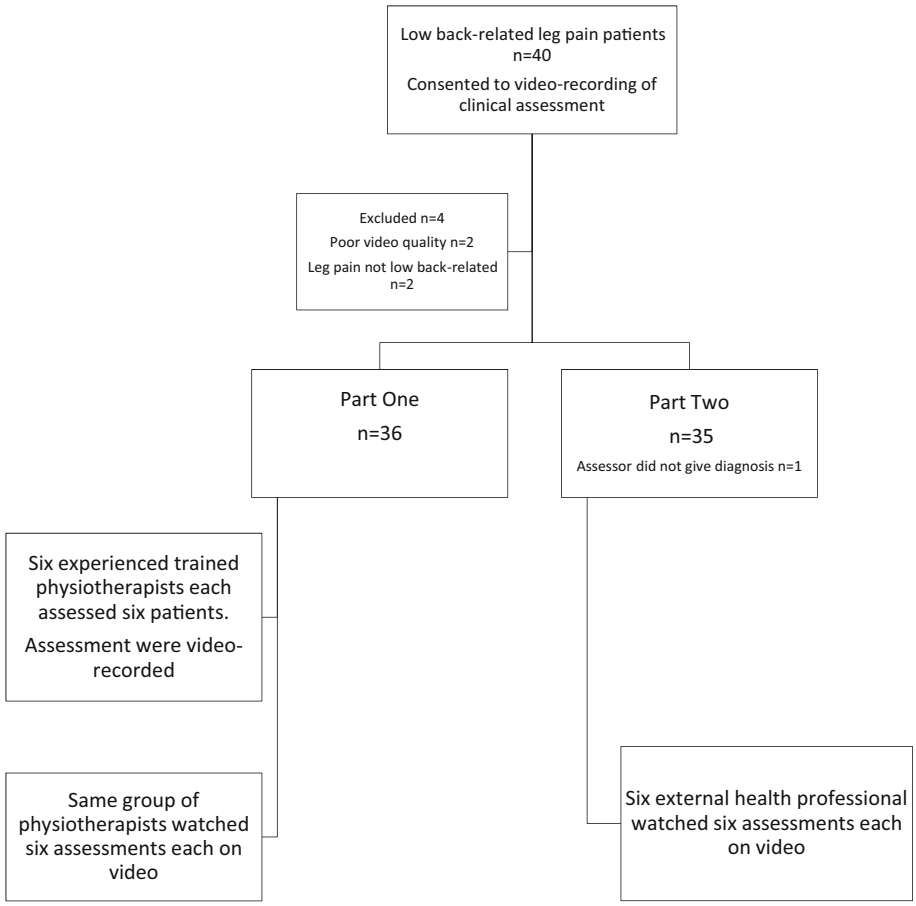

Fig. 1 Reliability study flow chart

Results

The median age of the 36 participating patients was 51 years and 61 % were female. Over half (58 %) had pain below the knee. A summary of the descriptive characteristics of the 36 patients is presented in Table 2. The six physiotherapists who performed the clinical assessments (Group A) and viewed the videos (Group B) were qualified on average 19.5 years (range 7–41 years) with an average of 15 years' experience (range 6–27 years) in predominately treating musculoskeletal conditions. The six health professionals who also viewed the videos (Group C) included two physiotherapists, a specialist registrar in rheumatology, a GP, a chiropractor and an osteopath. They were qualified on average for 20 years (range 14–26 years) and the allied health professionals had an average of 20.5 years' experience (range 15–26 years) in predominately treating musculoskeletal patients.

In part one, the physiotherapists diagnosed NRI in 25 of the 36 patients (Table 3) when assessing and watching the videos. Overall observed agreement was 72 % (expected agreement 58 %) with a kappa of 0.35 (95 % CI 0.02, 0.68) which is considered "fair" reliability [19]. In part two, between the physiotherapist who did the assessment and the health practitioners who watched the assessments on video, observed agreement was 71 % (expected agreement 57 %) with a kappa of 0.34 (95 % CI 0.02, 0.69) almost identical to results from part one (Table 4).

Table 3 Frequencies of patients ($n = 36$) classified by the physiotherapists raters in Group A and Group B as having either NRI or referred leg pain (part one)

	Raters in Group A		
	NRI	Referred	Total
Raters in Group B			
NRI	20	5	25
Referred	5	6	11
Total	25	11	36

Table 4 Frequencies of patients ($n = 35$) classified by the physiotherapists raters in Group A and health professionals in Group C as having either NRI or referred leg pain (part two)

	Raters in Group A		
	NRI	Referred	Total
Raters in Group C			
NRI	19	4	23
Referred	6	6	12
Total	25	10	35

Confidence in diagnosis

Agreement and reliability indices were calculated for levels of confidence in diagnosis (range 55–95 %). A clear and almost identical trend was seen in both part one and part two with agreement and the kappa coefficient increasing as confidence in diagnosis increased (Fig. 2). This trend of

Table 2 Descriptive characteristics of sample

Study sample $n = 36$		
Sex		
Male	14 (39 %)	
Female	22 (61 %)	
Age (years) median (range)	51 (23–74)	
Intensity back pain[a] (0–10) mean [standard deviation (SD)]	5.6 (2.8)	
Intensity leg pain[a] (0–10) mean (SD)	5.3 (2.7)	
RMDQ disability score[b] (0–23)	13.4 (6.2)	
Duration of pain	Back	Leg
0–6 weeks	13 (36 %)	16 (44 %)
6–12 weeks	10 (28 %)	8 (22 %)
>3 months	13 (36 %)	13 (34 %)
Pain below knees	21 (58 %)	
Off work because of back/leg pain	4 (11 %)	
Reduced hours/duties	3 (8 %)	

[a] Pain intensity measured using the mean of three 0–10 numerical rating scales for least and usual back pain over the previous 2 weeks and current back pain intensity (Dunn et al. [20])

[b] Roland Morris Disability Questionnaire leg pain version with scores from 0 to 23 with higher scores indicating higher disability (Patrick et al. [21])

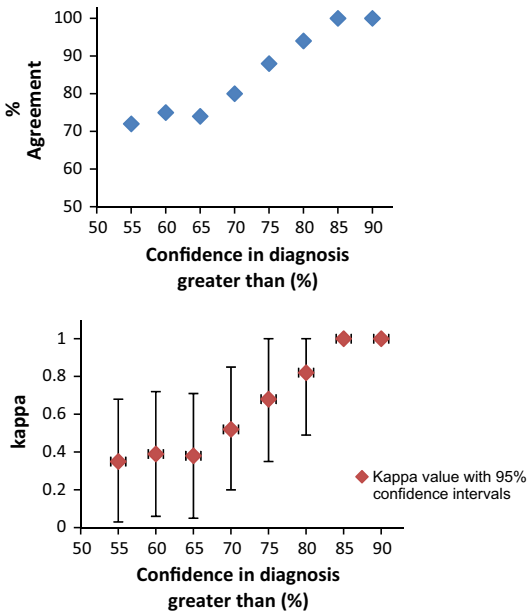

Fig. 2 Effect of increasing confidence in diagnosis on agreement and kappa coefficient (part one)

increasing agreement and reliability indices was noticeably evident once confidence in diagnosis was greater than 70 %. In the eighteen cases where both raters were over 80 % confident in their diagnosis, kappa was 0.82, considered almost perfect agreement [19].

The physiotherapists performing the assessment had the same levels of confidence in their diagnosis as when they watched the assessments on video (85 % median percentage confidence). Diagnostic confidence of raters in group C was slightly lower at 80 %. Median confidence in diagnosis for all raters was higher in cases of agreement (90 %) compared to cases of disagreement (70 %). There were ten disagreement cases in both part one and part two. Nine of the ten disagreement cases were the same for both parts.

Discussion

In this study, we have shown that the reliability of diagnosing nerve root involvement in LBLP patients with symptoms of any duration and severity is fair among experienced clinicians. Percentage agreement for both parts of the study and reliability as measured by the kappa coefficient were 72 and 71 % and 0.35 and 0.34, respectively. The agreement percentage is reasonable but kappa values under 0.6 are considered below the minimum standards for reliability coefficients [15]. The range of diagnostic confidence in this study varied between 50 and 100 % and further analysis showed that when both raters' confidence in clinical diagnosis was higher (over 70 %, $n = 30$), levels of agreement and reliability improved substantially (as shown in Fig. 2).

Numerous studies have reported on reliability of multi category classification systems for LBP. These systems are based on specific algorithms which possibly make it easier to agree on categories [12, 22]. This study reflects current clinical practice where an overall clinical impression is made based on the signs and symptoms. One other study looked specifically at the reliability of the overall clinical impression when assessing LBLP patients [13]. Reliability was substantial (kappa of 0.66) among pairs of neurologists who consecutively examined 91 patients with a new episode of sciatica "of sufficient intensity to justify 14 days of bed rest". However, comparing kappa values between studies is considered limited due to the differences in methods and sample characteristics [13, 23]. One explanation for the low kappa value seen in this reported study is that subjects were an unselected group, recruited from primary care with symptoms of varying degrees of severity and duration. The greater the proportion of patients with very clear symptoms or findings indicative of the condition of interest, the easier it is for different observers to agree [13] and conversely agreement on diagnosis may decrease with a greater proportion of "difficult to decide on" patients [24]. This was reflected in this study by the levels of confidence in diagnosis. Confidence was lower in cases of disagreement and higher levels of agreement and reliability were seen when diagnostic confidence increased.

The differing interpretations of clinical signs and symptoms among raters may also explain the kappa values. Despite consensus that a comprehensive clinical assessment is the cornerstones to a sound diagnostic process for LBLP [2, 25] inconsistencies are evident in studies when it comes to defining the specific criteria for diagnosing NRI [26]. Although diagnostic accuracy of individual items in clinical assessment of NRI is poor [5, 6] clinicians most likely give more weight to certain positive signs when making a confident diagnosis. To improve reliability of this study, fulfilling predefined criteria to make a NRI diagnosis as opposed to giving an overall clinical impression could have been specified. However, as highlighted above, as of yet, clear diagnostic criteria for confidently identifying NRI have not been agreed on.

Training of assessors and standardisation of procedures aim to minimize bias in reliability studies [15]. This study sought a balance between an appropriate level of standardisation and a setting that reflects current practice in primary care. Using multiple pairs of raters enhances generalizability and reduces the effect of rater bias. Although the physiotherapists were all experienced senior clinicians, very similar results were seen among clinicians from varied backgrounds. Regardless of training, standardisation or professional background, reliability was

merely fair when diagnosing LBLP, indicating that differentiating between some of these patients is a diagnostic challenge for clinicians in primary care. Irrespective of level of standardisation of procedures used, it is probably difficult to standardise the interpretation of a test result [3] and this is probably where most of the variation in clinical diagnosis comes from. Not all patients are difficult to diagnose, but this study showed that cases that are difficult to diagnose, contribute to reducing reliability indices among clinicians.

As yet we do not have universal agreement on criteria for differentiating between those patients who do and do not have NRI, or agreement on which combination of items from clinical assessment are more highly indicative of NRI. Diagnostic modelling in primary care LBLP populations, which assigns weights to various combinations of signs and symptoms has not been done at the point of writing, but could be a very helpful clinical diagnostic tool.

Strengths and limitations

Use of video lends strengths and limitations to the study design. It allows several raters to make independent diagnoses as opposed to burdening the patient with a repeated assessment and potentially aggravating their symptoms. Although video recording is considered an established method of recording GP consultations for research purposes [27] it has not been used in studies involving LBP patients that investigate the reliability of clinical diagnosis.

However, the use of video could lead to the Hawthorne effect i.e. that behaviour of patients or clinicians would alter due to being videoed, although a review of video recording in general practice found no conclusive evidence of the Hawthorne effect [27]. Physiotherapists performing the assessment had the same levels of confidence in their diagnosis as when they watched the assessments on video, possibly indicating that their performance and decision making were not influenced by being video recorded. The two groups of raters who watched the assessments on video did make very similar diagnostic decisions. In the case of diagnostic disagreement it is not possible to know whether the method of watching a video of a clinical examination negatively influences the ability to interpret the results of a test which contributes to diagnostic decisions. The researcher was present for all the viewings of the videos by raters in Group B and was rarely asked to clarify outcomes of tests. Raters in Group C watched the videos in their own home or work and did not contact the researcher to discuss any of the video assessments. The non-standardisation of the method of video watching makes it difficult to draw conclusions about its robustness as a test–retest method for diagnostic decision making.

The study sample represented patients from primary care seen in daily clinical practice. The number of patients recruited in this reliability study is similar to the majority of published reliability studies on LBP classification systems [12, 22]. However, the sample size calculation is based on specifying a zero value for kappa in the null hypothesis. The null hypothesis should ideally be set at a higher level, usually $\geq$0.4 which is considered more clinically acceptable [17]. However, to use this higher kappa cutoff as the sample size requirement, would require a sample size of 255 subjects [17] which was not practically feasible.

Conclusion

In this study, clinicians demonstrated different overall diagnostic impressions following assessments of LBLP patients which led to a fair reliability rating on their diagnostic decision. Some of this variability may have come from the methodology of using video recording but the diversity of signs and symptoms that these patients present with and the lack of clear guidelines as to what are the strongest criteria for differentiating between NRI and referred leg pain cannot be ignored. Ways of improving clinician agreement on diagnosis requires further exploration and one solution may be to assist the diagnosis process by identifying the optimal combination of items from the clinical assessment that best discriminate between these patients.

Acknowledgments This paper presents independent research funded by the National Institute for Health Research (NIHR). S Stynes is funded by a NIHR/CNO Clinical Doctoral Research Fellowship (CDRF-2010-055). Dr Konstantinou is funded by an HEFCE/NIHR Senior Clinical Lectureship. Professor Hay is a NIHR Senior Investigator. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Our thanks are expressed to all the raters who took part in this study: Corinna Dennison, Carol Doyle, Lucy Huckfield, Kika Konstantinou, Alan Nagington, Karen Williams, John Edwards, David Evans, Jonathan Field, Zoe Paskins, Gillian Pink, Paula Salmon. We also thank the patients who participated in this study.

Compliance with ethical standards

Conflict of interest The authors have no conflict of interests to declare.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Haswell K, Gilmour J, Moore B (2008) Clinical decision rules for identification of low back pain patients with neurologic involvement in primary care. Spine 33(1):68–73

2. Bogduk N (2009) On the definitions and physiology of back pain, referred pain, and radicular pain. Pain 147(1–3):17–19
3. Mannion A, Mutter U, Fekete T, Porchet F, Jeszenszky D, Kleinstuck F (2014) Validity of a single-item measure to assess leg or back pain as the predominant symptom in patients with degenerative disorders of the lumbar spine. Eur Spine J 23(4):882–887
4. Scholz J, Mannion RJ, Hord DE, Griffin RS, Rawal B, Zheng H, Scoffings D, Phillips A, Guo J, Laing RJC, Abdi S, Decosterd I, Woolf CJ (2009) A novel tool for the assessment of pain: validation in low back pain. PLoS Med 6(4):e1000047
5. Vroomen P, de Krom MC, Knottnerus J (1999) Diagnostic value of history and physical examination in patients with sciatica due to disc herniation; a systematic review. J Neurol 246(10):899–906
6. van der Windt DA, Simons E, Riphagen I, Ammendolia C, Verhagen AP, Laslett M, Deville W, Deyo RA, Bouter LM, De Vet HCW, Aertgeerts B (2010) Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain. Cochrane Database Syst Rev 2(2):CD007431
7. Jensen M, Brant-Zawadski M, Obuchowski N, Modic M, Malkasian D, Ross J (1994) MRI imaging of the lumbar spine in people without back pain. New Engl J Med 331(2):69–73
8. Iversen T, Solberg TK, Romner B, Wilsgaard T, Nygaard O, Waterloo K, Brox JI, Ingebrigtsen T (2013) Accuracy of physical examination for chronic lumbar radiculopathy. BMC Musculoskelet Disord 14:206. doi:10.1186/1471-2474-14-206
9. Karppinen J, Malmivaara A, Tervonen O, Pääkkö E, Kurunlahti M, Syrjälä P, Vasari P, Vanharanta H (2001) Severity of symptoms and signs in relation to Magnetic Resonance Imaging findings among sciatic patients. Spine 26(7):E149–E154
10. Coggon D, Martyn C, Palmer K, Evanoff B (2005) Assessing case definitions in the absence of a diagnostic gold standard. Int J Epidemiol 34(4):949–952
11. McCarthy CJ, Gittins M, Roberts C, Oldham JA (2007) The reliability of the clinical tests and questions recommended in international guidelines for low back pain. Spine 32(8):921–926
12. Smart K, Curley A, Blake C, Staines A, Doody C (2010) The reliability of clinical judgments and criteria associated with mechanisms-based classifications of pain in patients with low back pain disorders: a preliminary reliability study. J Manual Manip Therapy 18(2):102–110
13. Vroomen P, de Krom MC, Knottnerus J (2000) Consistency of history taking and physical examination in patients with suspected lumbar nerve root involvement. Spine 25(1):91–96
14. Busse J, Riva J, Nash J, Hsu S, Fisher C, Wai E, Brunarski D (2013) Surgeon attitudes toward nonphysician screening of low back or low back-related leg pain patients referred for surgical assessment: a survey of Canadian spine surgeons. Spine 38(7):E402–E408
15. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, Roberts C, Shoukri M, Streiner DL (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 64(1):96–106
16. Konstantinou K, Beardmore R, Dunn KM, Lewis M, Hider SL, Sanders T, Jowett S, Somerville S, Stynes S, van der Windt DA, Vogel S, Hay EM (2012) Clinical course, characteristics and prognostic indicators in patients presenting with back and leg pain in primary care, The ATLAS study. BMC Musculoskelet Disord 13:4. doi:10.1186/1471-2474-13-4
17. Sim J, Wright C (2005) The kappa statistic in reliability studies: use, interpretation and sample size requirements. Phys Ther 85(3):257–268
18. Konstantinou K, Hider S, Vogel S, Beardmore R, Somerville S (2012) Development of an assessment schedule for patients with low back-associated leg pain in primary care: a Delphi consensus study. Eur Spine J 21(7):1241–1249
19. Landis JR, Koch GC (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
20. Dunn KM, Jordan KP, Croft PR (2010) Recall of medication use, self-care activities and pain intensity: a comparison of daily diaries and self-report questionnaires among low back pain patients. Prim Health Care Res Dev 11:93–102
21. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB (1995) Assessing health-related quality of life in patients with sciatica. Spine 20(17):1899–1908
22. Schafer A, Hall TM, Ludtke K, Mallwitz J, Briffa N (2009) Interrater reliability of a new classification system for patients with neural low back-related leg pain. The Journal of Manual and Manipulative Therapy 17(2):109–117
23. Byrt T, Bishop J, Carlin J (1993) Bias, prevalence and kappa. J Clin Epidemiol 46(5):423–429
24. Vach W (2005) The dependence of Cohen's kappa on the prevalence does not matter. J Clin Epidemiol 58(7):655–661
25. Freynhagen R, Rolke R, Baron R, Tolle TR, Rutjes A-K, Schu S, Treede R-D (2008) Pseudoradicular and radicular low-back pain—a disease continuum rather than different entities? Answers from quantitative sensory testing. Pain 135(1–2):65–74
26. Lin C, Verwoerd A, Maher C, Verhagen A, Pinto R, Luijsterburg P, Hancock M (2014) How is radiating leg pain defined in randomized controlled trials of conservative treatments in primary care? A systematic review. Eur J Pain 18(4):455–464
27. Coleman T (2000) Using video-recorded consultations for research in primary care: advantages and limitations. Fam Pract 17(5):422–427