

Bo C. Bertilson
Johan Bring
Anneli Sjöblom
Karin Sundell
Lars-Erik Strender

Inter-examiner reliability in the assessment of low back pain (LBP) using the Kirkaldy-Willis classification (KWC)

Received: 12 November 2004
Revised: 5 October 2005
Accepted: 1 December 2005
Published online: 25 January 2006
© Springer-Verlag 2006

B. C. Bertilson (✉) · L.-E. Strender
Karolinska Institutet, Center for Family
Medicine, Alfred Nobels ellé 12,
Huddinge, Stockholm 141 83, Sweden
E-mail: bo.bertilson@ki.se
Tel.: +46-70-320232
Fax: +46-8-52488707

J. Bring
Statistics, Uppsala University,
Uppsala, Sweden

A. Sjöblom · K. Sundell
Älvsjö Rygg- och Sportrehab,
Stockholm, Sweden

Abstract Reliable classification systems and clinical tests are sought for the care of patients with low back pain (LBP). The objectives of this clinical study were to evaluate inter-examiner reliability in the classification of patients with LBP, the influence of radiological findings on the classification and the reliability of some clinical tests. Two examiners independently assessed 50 outpatients with LBP. Inter-examiner reliability in classification of patients with LBP using Kirkaldy-Willis classification (KWC) system and in 30 clinical tests was calculated as percentage agreement and kappa coefficients (κ). Inter-examiner reliability was excellent ($\kappa > 0.8$) for classification according to KWC. Radiological findings did not influence the reliability. Age of the

patient, movement range, and pain and neurological signs seemed to guide the decision on classification. The reliability of clinical tests was good ($\kappa > 0.6$) in 6 tests and moderate ($\kappa > 0.4$) in 12 tests. Good inter-examiner reliability was found for the SLR test, movement range and sensibility testing with spurs in dermatome areas. We conclude that the KWC for classifying patients with LBP seems to be a reliable classification system depending on a few key observations and that moderate and good inter-examiner reliability can be achieved in several clinical tests in the assessment of LBP.

Keywords Low back pain · Classification · Inter-examiner reliability · Assessment of clinical tests · Kirkaldy-Willis

Introduction

The lack of a practical and reliable classification system for low back pain (LBP) syndromes is considered a top-priority research question as these syndromes have a substantial influence on health and quality of life, and impose enormous costs on health systems [7, 16, 18, 24]. Studies on classification systems with a pathoanatomical viewpoint like the Kirkaldy-Willis Classification (KWC) are rare [32]. Our hypothesis in this clinical study on outpatients with LBP is that KWC may be a reliable classification system.

Riddle in a review article on classification of LBP notes that “the most compelling argument for developing and using classification systems is that our current

system for grouping patients appear to be inadequate” [35]. This inadequacy seems clear as only 15% of LBP cases are said to have a specific diagnosis [32, 35]. The rest are classified as “non-specific” LBP, where psychosocial factors are often mentioned as being causative [21, 28].

Then, what motives can be listed for classifying LBP? Sahrman gives an answer, “A critical step...is the development of diagnostic categories...of the signs and symptoms that are identified by ...examinations and tests. A primary premise is that treatment should be based on the diagnosis ...these diagnoses will (1) clarify practice, (2) provide an important means of communication with colleagues and consumers, (3) ...direct research and assessment of treatment effectiveness, and (4)

reduce the tendency toward cultism associated with practice based almost entirely on treatment approaches" [37]. We believe the lack of an evidence-based consensus on how to assess and classify LBP contributes to the fact that almost none of the commonly occurring and frequently practiced medical interventions for LBP have proven to be effective [17].

Previous studies on classification of LBP have focused on systems based on the effect of treatment [11, 13, 38, 42]. This focus is due to the assumption that a pathoanatomical etiology can rarely be determined. However, Bernard, Kirkaldy-Willis and others contend that a pathoanatomical viewpoint is fundamental to interpretation and treatment of LBP [3, 23, 27, 32].

The KWC of LBP identifies three phases of progressive degeneration of the disc(s) and/or facet joint(s) in the lower spine as a result of minor or major trauma [22]. In phase I, called "dysfunction", the typical patient is a person below middle age with acute/subacute LBP presenting after rotational or compressive trauma to the back resulting in capsular/annular tears and minor facet joint subluxation which may lead to synovitis. Sustained segmental muscle hypertonicity cause ischemia and altered metabolism in an injured segment which may lead to fibrosis, osteophyte formation and initiate disc degeneration. Phase I is considered the most common form of LBP. In phase II, called "unstable", the typical patient would be a middle age person with recurrent LBP where successive trauma to disc and facet joint(s) has caused degeneration of cartilage, laxity of the facet capsule and further tears and internal disruption of the disc which may result in bulging of the annulus. In phase

III, called "stabilisation", the typical person would be older and have a long history of LBP but now mainly distal symptoms. The pathoanatomical changes include destruction of cartilage and discs with fibrosis and osteophyte formation causing stiffness and risk for nerve impingement in the spine.

Progressive degeneration due to damage to the disc, endplate and/or adjacent joints and ligaments, has been noted by others as also the fact that nerve irritation can occur in any stage of disc degeneration due to chemical and/or mechanical stimuli [1, 5, 20, 30, 31]. We consider the KWC to be logical and practical as it takes into account key observations of symptoms, signs and radiological changes in the three KWC phases described in Table 1.

Our objectives with this study were; first to evaluate the inter-examiner reliability in classification of patients with LBP using the KWC system, second to evaluate if knowledge of radiological findings influence the classification and third to evaluate the inter-examiner reliability of some clinical tests in the assessment of LBP [39].

Materials and methods

Examiners

Two female physiotherapists (A and B), both with international certification in orthopaedic manual therapy and with 20, respectively, 24 years of clinical experience and 9 years of work at the same private outpatient clinic in southern Stockholm.

Table 1 Key observations in the phases of the KWC system of low back degenerative disease

Phase	I Dysfunction	II Unstable	III Stabilisation
Symptoms	Low back pain Often localised Sometimes referred Movement painful	Those of dysfunction Giving way of back: "catch" Pain on coming to standing Position after flexion	Less low back pain Mainly leg pain
Signs	Local tenderness Muscle contracted Hypomobility Extension painful Seldom neurology	Detection of abnormal movement (Inspection, palpation) Observation of "catch", sway, or shift When coming erect after Flexion	Muscle tenderness Stiffness Reduced movement Scoliosis
Radiological changes	Abnormal decreased movement Spinous processes malaligned Irregular facets Early disc changes	Anteroposterior Lateral shift Rotation Abnormal tilt Malaligned spinous processes Oblique Opening facets lateral Spondylolisthesis (in flexion) Retrospondylolisthesis (in extension) Abnormal opening of disc Abrupt change in pedicle height CT changes disc bulging	Some neurology Enlarged facets Loss of disc height Osteophytes Small foramina Reduced movement Scoliosis

Patients and randomisation

During 4 months, 50 patients visiting a private outpatient back clinic were independently assessed by each examiner. Physicians, mainly general practitioners, referred about 15% of the patients, the others came without referral. The inclusion criteria were LBP with or without radiation to the leg. Exclusion criteria were age below 16, previous spinal surgery or visit to the clinic or difficulty to perform a full exam.

Suitable patients were informed about and invited to participate in the study when they called to make their first appointment. For those interested, one of the 60 sealed envelopes, prepared by a statistician, was opened in order to evenly assign the patient to the first assessment with either examiner A or B.

No patient declined the offer to participate. Three patients did not show up at the appointed time and two patients were excluded, one due to obesity and the other due to confounding neck problems made known prior to entering the study. The characteristics of the 50 patients included are shown in Table 2. The regional ethics committee approved the study. Written information about the study was given as patients arrived for their first visit.

Procedure

Each assessment session consisted of history taking including information on radiological observations and then a structured physical examination before a decision on diagnosis and treatment recommendations was made by the respective examiner. An average of 40 min/patient was allocated for each examiner. The interval between assessments was less than 30 min. Standardised assessment forms were filled out for each patient where results of each clinical test and the KWC were noted. After each session the assessment form was placed in an envelope, which was then sealed. During the assessment sessions, the patient was not informed about the results of the examination.

Radiological examination (RE)

Some time prior to entering this study, 20 patients had a RE of their lower back. Of these 20, 15 had a plain

X-ray and 5 had an MRI. The radiologist's written assessment of the RE was available to examiners during the history taking. This was done in order to analyse these patients as a subgroup to find out whether the results of the RE influenced the KWC.

Physical examination

The physical examination included 30 clinical tests. The clinical test procedures and definitions on how to assess observations are presented in the Appendix. Assessments were made subjectively without a measuring device to resemble everyday practice. In all tests except one a binary (normal/not normal) decision was made. The exception was the evaluation of inter-segmental mobility, where the examiner had to decide whether mobility was decreased, normal or increased. A "not normal" decision was termed a "positive" observation. The tests included were chosen based on the examiners experience of assessing patients according to the KWC and had been used in the everyday practice of both examiners and also in our previous study on reliability of clinical tests [39].

Classification

For each patient, the examiners had to decide which of the three phases (I-III) in the KWC system (Table 1) the patient was in. Furthermore, when radiating pain was present the examiner had to decide whether there was nerve root involvement or not.

Statistics

The prevalence of positive observations and the degree of overall agreement was calculated in percent. The kappa coefficient (κ) was calculated to express inter-examiner reliability [10, 12]. Weighted κ values were calculated for the inter-segmental mobility tests and the KWC, where more than two options were possible [2]. For the inter-segmental mobility tests κ values were also calculated after observations were dichotomised, i.e. the two possible abnormal observations (increased or decreased) were combined to obtain a binary item (normal or not normal). For tests measured on the right and the left side, the prevalence of a positive test on one side or the other was used to indicate that the test was positive.

The κ is a chance-corrected measure of agreement that is influenced by the prevalence of positive findings and is attenuated most severely towards low values when the prevalence is either particularly low or high [41]. Therefore, the κ was not calculated when the mean of the examiners' prevalence was below 10% or above

Table 2 Patient characteristics for the entire sample ($n = 50$)

	Men	Women
Number	18	32
Mean age (range), years	34 (16-61)	38 (18-61)
Mean height (range), cm	181 (170-186)	167 (158-175)
Mean weight (range), kg	83 (70-95)	62 (49-77)

Table 3 Inter-examiner reliability in classifying LBP using the KWC ($n=50$)

Phase	Examiner A			Total #
	I Dysfunction	II Unstable	III Stabilisation	
Examiner B				
Dysfunction	21	1	1	23
Unstable	2	16	0	18
Stabilisation	1	1	7	9
Total	24	18	8	50

Boldface figures indicate total agreement on the classification, a total of 44

90%, or when the prevalence of one examiner was 0%. The standard error (SE) and the 95% confidence interval (CI) for the κ were also calculated. The κ was classified as follows: <0 "no agreement better than chance", 0–0.2 "poor", 0.21–0.4 "slight", 0.41–0.6 "moderate", 0.61–0.8 "good" and 0.81–1 "excellent" agreement [2, 21, 39].

The mean prevalence of positive clinical test observations for the two examiners was calculated for each diagnostic phase. Fisher's exact test was used to assess

the difference in likelihood of a positive observation in clinical tests in the three phases. The test was done for each examiner separately. The significance level $P < 0.05$ was used.

Results

Inter-examiner reliability in classifying patients with LBP using the KWC

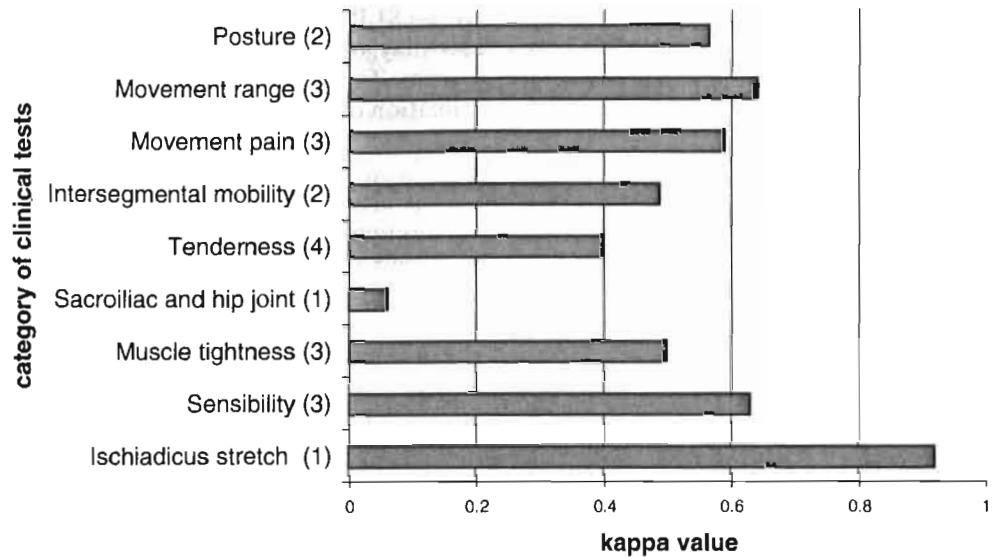
In 44 (88%) of the patients the examiners agreed on the KWC (Table 3). The weighted κ was 0.81 ± 0.105 (SE), indicating excellent reliability. Dysfunction was the most common phase. The characteristics for the 44 patients where examiners agreed fully on the KWC showed that mean age increased gradually from 29 years in phase I (range 16–43), to 35 years in phase II (range 22–54) and 56 years in phase III (range 46–61). The proportion of women to men was high in phases I (17 to 4) and III (5 to 2), whereas men were predominant in phase II (5 to 11). Nerve root involvement was present in 10 patients (20%) according to both examiners. The agreement was 100%, the κ was 1.0 ± 0.236 (SE).

Table 4 Inter-examiner reliability in the assessment of clinical tests ($n=50$)

Clinical test ^a	Prevalence of positive observations (%) by examiner		Overall agreement (%)	Kappa coefficient	SE for kappa	95% CI for kappa	
	A	B				Lower	Upper
	Posture						
Sagittal configuration	46	56	78	0.56	0.139	0.29	0.84
Movement range							
Flexion range	30	28	94	0.85	0.141	0.58	1.00
Extension range	26	32	78	0.47	0.140	0.19	0.74
Lateral bend range	14	22	88	0.60	0.136	0.33	0.86
Movement pain							
Flexion pain	50	44	86	0.72	0.140	0.44	1.00
Extension pain	34	40	78	0.53	0.140	0.25	0.80
Lateral bend pain	32	38	78	0.52	0.140	0.24	0.79
Inter-segmental mobility							
Segment above lumbosacral	46	40	74	0.47	0.140	0.20	0.75
Lumbosacral segment	82	84	86	0.50	0.141	0.23	0.78
Tenderness							
Springing test	62	64	74	0.44	0.141	0.17	0.72
Segment above lumbosacral	66	60	74	0.44	0.140	0.17	0.72
Lumbosacral segment	80	76	84	0.53	0.141	0.26	0.81
Muscle stiffness							
Rectus	20	16	88	0.59	0.140	0.32	0.87
Hamstring	24	34	82	0.57	0.137	0.30	0.84
Neurological							
Sensibility to pain L4	6	16	90	0.50	0.123	0.26	0.74
Sensibility to pain L5	12	20	92	0.71	0.135	0.44	0.97
Sensibility to pain S1	14	24	90	0.68	0.134	0.42	0.94
Ischiadicus stretch (SLR)	16	14	98	0.92	0.141	0.65	1.00

^aOnly clinical tests with $\kappa > 0.40$ are shown

Fig. 1 Mean kappa values for categories of clinical tests. The number of clinical tests with kappa coefficients included in the category are noted in parentheses (*n*).



Inter-examiner reliability in the assessment of clinical tests

In Table 4 are shown those 18 tests with $\kappa > 0.40$ indicating at least moderate reliability. The highest κ were found for the SLR test (0.92), the flexion pain (0.72) and sensibility tests for L5 (0.71) and S1 (0.68). Three tests had $\kappa < 0.40$; the sacroiliac compression pain (0.06), paravertebral tenderness (0.17) and iliopsoas tightness (0.33). The weighted κ for inter-segmental mobility was 0.57 ± 0.106 (SE) for the lumbosacral segment and 0.45 ± 0.104 (SE) for the previous segment (not shown in any table). Taking into account the lower bound of the 95% CI for κ , 16 of the 22 clinical tests where κ could be calculated receive $\kappa < 0.40$. When clinical tests are grouped in categories (Fig. 1) Ischiadicus stretch (SLR), the movement range and the sensibility tests show the highest κ (> 0.6). The highest prevalence of positive observations, four out of five, was found for deranged inter-segmental mobility and tenderness in the lumbosacral segment.

Prevalence of positive observations in tests distinguishing the KWC phases

The prevalence of positive observations in tests distinguishing the KWC phases is shown in Table 5. There were six clinical tests where both examiners showed a significant difference in likelihood for a positive observation distinguishing KWC phases. Patients classified in phase II had a significantly higher likelihood for a positive test compared to patients in phase I and III for the following tests: Lordosis, Flexion range, Flexion pain, Strength big toe extension and Ischiadicus stretch

(SLR). For Extension pain, patients in phase III were significantly less likely to have a positive observation compared to patients in phase I and II.

Discussion

Our study showed that it was possible to attain excellent ($\kappa > 0.8$) inter-examiner reliability in classifying patients with LBP using the KWC. Moreover, that knowledge of RE did not improve the reliability. Rather, the age of the patient, movement range and pain and a few positive neurological observations seemed to have the major impact on the classification. Furthermore, we found that some clinical tests can be performed with moderate or

Table 5 Prevalence of positive observations in tests distinguishing the KWC phases

Clinical test ^a	Prevalence of positive observations (%) in each phase ^b		
	I Dysfunction	II Unstable	III Stabilisation
Lordosis	0	25	0
Flexion range	11	61	12
Flexion pain	32	78	24
Extension pain	40	50	0
Strength big toe extension	0	22	0
Ischiadicus stretch (SLR)	0	42	0

^aOnly those clinical tests where a significant difference in likelihood for both examiners of a positive observation in the KWC phases are shown, the other tests did not differ in likelihood between phases for both examiners

^bMean of the two examiners' observations

even good inter-examiner reliability, especially the SLR, movement range and the sensibility tests. This may be the first study on the reliability of the KWC system. The strength of the study is that we included an evaluation of the reliability of the clinical tests we used in the process of classification.

Earlier inter-examiner reliability studies on the classification of LBP have focused on treatment options. The McKenzie system, based on impairments, seems to be the most evaluated and inter-examiner reliability ranges from κ 0.26 to 0.70 in different studies [21, 34, 36]. Delitto et al. [11] have proposed a treatment-based classification system that in one study has shown moderate inter-examiner reliability. Wilson et al. [42] in 1999 presented a classification system consisting of five patterns of mechanical back pain where examiners using key elements of the history and the clinical examination agreed on patient classification in 79% of 204 patients ($\kappa=0.61$).

The Quebec Task Force, Wadell and Turk and BenDebba et al have proposed classifications systems based mainly on the spatial distribution of pain and/or the presence of signs of mechanical nerve root pain, but to our knowledge their reliability has not been studied [4, 25].

The excellent inter-examiner reliability of the KWC system in our study is equal to the best results of earlier studies on classification of LBP as aforementioned [13, 21]. Our examiner's experience in classifying according to KWC was long since established and part of their daily thinking. The question is—what observations may have led to this excellent result? Surprisingly, radiological observations did not improve the reliability. One reason may be that the examiners were physiotherapists and specially trained to make diagnoses based on history and clinical observations [8, 14, 19]. The observation having the greatest impact on the classification seems to be an increase in age in the subsequent phases of KWC. This observation is in accordance with the theory of progressive degeneration in phase I through III. The painful extension in phase I and II but not in phase III is also in accordance with the KWC key observations. The tendency of decreased inter-segmental mobility through phase I–III was logical, but not statistically significant. Other key observations described in the KWC were not verified in our study. For example, tenderness and muscle stiffness were signs almost equally frequent in patients of all three phases of the KWC. The presence of positive neurological observations in phase II is not a key sign noted by Kirkaldy-Willis. However, we consider the presence of neurological observations in phase II logical in accord with the theory of discoligamentous injuries that can exert chemical and/or mechanical injury to spinal nerve tissue [29–31].

The inter-examiner reliability of the clinical tests in this study confirms the results of a previous study of our

own [39]. The most reliable categories of clinical tests with $\kappa > 0.6$, indicating good reliability were the SLR, the movement range and the sensibility to pain tests (Table 4). Within these categories were five of the six tests that reached $\kappa > 0.6$ and also four of the five tests where the lower limit of the 95% CI was above $\kappa > 0.4$. The SLR test has shown good reliability in other studies [26, 39], although poor reliability has also been reported [40]. The pinwheel sensibility test is not well known, but it is easy to perform and has demonstrated good reliability in one of our studies on neck–shoulder patients [6]. This test may be recommended for further studies on reliability and validity. The least reliable test was the sacroiliac compression test. All other categories of tests were found with moderate reliability (Fig. 1).

The reliability of clinical tests may depend on several factors. First, how well a test can be standardised. Tenderness and sacroiliac compression tests are difficult to standardise and logically receives low kappa values in ours as in other studies [9, 26, 33, 39]. Second, it has been reported that reliability improves with an increased prevalence of positive observations [9]. However, in our study the tests where we had few positive observations (the SLR and the sensibility tests) had the highest kappa values, and so we question if prevalence is important to reliability. Third, the experience of the examiners may vary, but having experienced examiners may not be sufficient. One study even indicated decreased reliability with increased experienced examiners. A reason for this may be that experienced examiners develop personal idiosyncrasies [15]. We cannot exclude a shared bias on the part of our examiners neither in the classification of LBP nor in the assessment of clinical tests, and this limitation needs further studies to dismiss.

Other limitations to our study are that the patient sample is rather small, although not smaller than the late studies on the Mc Kenzie system [21, 34, 36]. Furthermore, an evaluation of correlation between radiological and clinical observations would have been valuable. This evaluation is now being done by our team in a larger study. Furthermore, all assessments were subjective and not statistical, as it has been suggested to be the ideal [35]. A statistical method of assessing history, clinical and radiological observations is yet to be developed and elaborated on in future studies. Future studies may also evaluate if the KWC can improve treatment selection and outcome.

Conclusions and clinical implications

Excellent inter-examiner reliability was found for the KWC system of LBP. In clinical practice, it appears that KWC could be based on knowledge of the age of the patient and a few clinical tests. KWC phase I would be a young patient with pain on back extension, yet normal

range of motion and no neurological signs. Phase 2 would be a middle age patient with pain on back flexion and extension, reduced flexion, altered lordosis, extensor hallucis longus weakness and a positive SLR. Phase 3 would be an older patient with no pain on movement and no neurological signs. Knowledge of RE is not necessary.

Good inter-examiner reliability was found for clinical tests like the SLR, movement range and sensibility testing with spurs in dermatome area. We conclude that the KWC and the clinical tests just mentioned may be recommended for use and further studies in the examination and classification of LBP.

Acknowledgement This study was supported by funds from Stockholm county council.

Appendix

Clinical test procedures and definitions on what was evaluated as not normal (NN)

First with patient standing

- Posture. Increased or decreased lumbar lordosis respectively scoliosis NN.
- Movement range. Decreased lumbar forward flexion, extension and lateral bending NN.
- Movement pain. Pain on lumbar flexion, extension and on lateral bending NN.
- Foramen compression test was performed with lateral bending and rotation of the lumbar spine to the tested side. Provoked pain radiating down below the knee NN.

Second with patient lying in prone position

- Femoral nerve stretch test (Ely's test). The tested leg was passively extended in the hip joint and flexed in the knee joint. Provoked pain radiating down to the anterior thigh NN if it could be increased by flexion of the head and/or plantar flexion in the ankle joint.
- Muscle stiffness of rectus femoris NN if the heel of the foot did not reach the gluteal skin.
- Springing test. One hand placed on fingers of the other hand positioned on each side of the spinal processes. Tenderness or decreased elasticity NN.
- Sacroiliac compression pain was evaluated as the lateral edge of the sacrum was compressed using both hands. Increased pain on either side NN.

- Paravertebral tenderness. Evaluated in the lumbar area between the spinal processes and the midaxillary line. Tenderness and/or a difference between left and right side NN.
- Inter-segmental tenderness in the lumbosacral segment and the segment immediately above. Tenderness on palpation with fingers NN.

Third with patient lying on one side with the hips and knees flexed

- Inter-segmental mobility (angular and translational) in the lumbosacral segment and the segment immediately above were evaluated by palpation while passively moving the patients knees and classified as decreased, normal or increased.

Fourth with patient lying in supine position

- Ischiadicus nerve stretch test (straight leg raising or SLR) and hamstring stiffness test was performed simultaneously. The examiner fixed one leg to the table to stabilise the pelvis, while elevating the tested leg with the knee in extension. SLR NN if pain radiated below the knee and if the pain increased either when the head was flexed or when the foot was dorsiflexed. Hamstring test NN if less than 80° flexion in the hip was reached.
- Sensibility was tested with a pinwheel, one side at a time, in the dermatome areas according to the maps of Netter. Deranged or asymmetrical sensibility NN.
- Strength in the ankle and in large toe dorsiflexion NN if decreased or asymmetrical.
- Patellar and Achilles reflexes NN if the response was deranged or asymmetrical.
- Internal hip rotation. Evaluated with the hip and knee in 90° flexion. Decreased range and/or asymmetry in the joint motion NN.
- Patrick's test. Performed as described by K. Lewit in *Manuelle Medizin* from Munchen-Wien-Baltimore, 1987. Pain in the dorsal region of the sacroiliac joint on the same side and/or decreased range of movement with an increased resistance at the end point NN.
- Iliopsoas muscles stiffness was tested with the tuber os ischi at the lower end of the examination table. To maintain the lower back flat on the table, the opposite leg was maximally flexed in the hip joint and held by the patient against the chest. If the thigh on the side tested did not reach the horizontal plane of the examination table NN.

References

1. Adams MA, Freeman BJ, Morrison HP, Nelson IW, Dolan P (2000) Mechanical initiation of intervertebral disc degeneration. *Spine* 25:1625-1636
2. Altman (1999) Practical statistics for Medical Research
3. Banic B, Petersen-Felix S, Andersen OK, Radanov BP, Villiger PM, Arendt-Nielsen L, Curatolo M (2004) Evidence for spinal cord hypersensitivity in chronic pain after whiplash injury and in fibromyalgia. *Pain* 107:7-15
4. BenDebba M, Torgerson WS, Long DM (2000) A validated, practical classification procedure for many persistent low back pain patients. *Pain* 87:89-97
5. Benneker LM, Heini PF, Anderson SE, Alini M, Ito K (2005) Correlation of radiographic and MRI parameters to morphological and biochemical assessment of intervertebral disc degeneration. *Eur Spine J* 14:27-35
6. Bertilson BC, Grunnesjo M, Strender LE (2003) Reliability of clinical tests in the assessment of patients with neck/shoulder problems-impact of history. *Spine* 28:2222-2231
7. Borkan J (1998) A report from the second international forum for primary care research on low back pain; reexamining priorities. *Spine* 23(18):1992-1996
8. Brant-Zawadzki MN, Jensen MC, Obuchowski N, Ross JS, Modic MT (1995) Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities. A comparison of two nomenclatures. *Spine* 20:1257-1263 discussion 1264
9. Carmichael JP (1987) Inter- and intra-examiner reliability of palpation for sacroiliac joint dysfunction. *J Manipulative Physiol Ther* 10:164-171
10. Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 43:551-558
11. Delitto A, Erhard RE, Bowling RW (1995) A treatment-based classification approach to low back syndrome: identifying and staging patients for conservative treatment. *Phys Ther* 75:470-485 discussion 485-479
12. Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 43:543-549
13. Fritz JM, George S (2000) The use of a classification approach to identify subgroups of patients with acute low back pain. Interrater reliability and short-term treatment outcomes. *Spine* 25:106-114
14. Fritz JM, Wainner RS, Hicks GE (2000) The use of nonorganic signs and symptoms as a screening tool for return-to-work in patients with acute low back pain. *Spine* 25:1925-1931
15. Gonnella C, Paris SV, Kutner M (1982) Reliability in evaluating passive intervertebral motion. *Phys Ther* 62:436-444
16. Greenough CG, Fraser RD (1992) Assessment of outcome in patients with low-back pain. *Spine* 17:36-41
17. Hansson TH, Hansson EK (2000) The effects of common medical interventions on pain, back function, and work resumption in patients with chronic low back pain: a prospective 2-year cohort study in six countries. *Spine* 25:3055-3064
18. Jensen IB, Bodin L, Ljungqvist T, Gunnar Bergstrom K, Nygren A (2000) Assessing the needs of patients in pain: a matter of opinion? *Spine* 25:2816-2823
19. Jensen MC, Brant-Zawadzki MN, Obuchowski N, Modic MT, Malkasian D, Ross JS (1994) Magnetic resonance imaging of the lumbar spine in people without back pain. *N Engl J Med* 331:69-73
20. Kayama S, Olmarker K, Larsson K, Sjogren-Jansson E, Lindahl A, Rydevik B (1998) Cultured, autologous nucleus pulposus cells induce functional changes in spinal nerve roots. *Spine* 23:2155-2158
21. Kilpikoski S, Airaksinen O, Kankaanpaa M, Leminen P, Videman T, Alen M (2002) Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine* 27:E207-214
22. Kirkaldy-Willis WH (1999) Managing low back pain
23. Kirkaldy-Willis WH, Hill RJ (1979) A more precise diagnosis for low-back pain. *Spine* 4:102-109
24. Koes BW, van Tulder MW, Ostelo R, Kim Burton A, Waddell G (2001) Clinical guidelines for the management of low back pain in primary care: an international comparison. *Spine* 26:2504-2513 discussion 2513-2504
25. Loisel P, Vachon B, Lemaire J, Durand MJ, Poitras S, Stock S, Tremblay C (2002) Discriminative and predictive validity assessment of the quebec task force classification. *Spine* 27:851-857
26. McCombe PF, Fairbank JC, Cockersole BC, Pynsent PB (1989) Volvo Award in clinical sciences. Reproducibility of physical signs in low-back pain. *Spine* 14:908-918
27. McGill SM (1998) Low back exercises: evidence for improving exercise regimens. *Phys Ther* 78:754-765
28. Michel A, Kohlmann T, Raspe H (1997) The association between clinical findings on physical examination and self-reported severity in back pain. Results of a population-based study. *Spine* 22:296-303 discussion 303-294
29. Olmarker K, Holm S, Rosenqvist AL, Rydevik B (1991) Experimental nerve root compression. A model of acute, graded compression of the porcine cauda equina and an analysis of neural and vascular anatomy. *Spine* 16:61-69
30. Olmarker K, Nordborg C, Larsson K, Rydevik B (1996) Ultrastructural changes in spinal nerve roots induced by autologous nucleus pulposus. *Spine* 21:411-414
31. Olmarker K, Rydevik B, Nordborg C (1993) Autologous nucleus pulposus induces neurophysiologic and histologic changes in porcine cauda equina nerve roots. *Spine* 18:1425-1432
32. Petersen T, Olsen S, Laslett M, Thorsen H, Manniche C, Ekdahl C, Jacobsen S (2004) Inter-tester reliability of a new diagnostic classification system for patients with non-specific low back pain. *Aust J Physiother* 50:85-94
33. Potter NA, Rothstein JM (1985) Inter-tester reliability for selected clinical tests of the sacroiliac joint. *Phys Ther* 65:1671-1675
34. Razmjou H, Kramer JF, Yamada R (2000) Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low-back pain [In Process Citation]. *J Orthop Sports Phys Ther* 30:368-383 discussion 384-369
35. Riddle DL (1998) Classification and low back pain: a review of the literature and critical analysis of selected systems. *Phys Ther* 78:708-737
36. Riddle DL, Rothstein JM (1993) Inter-tester reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine* 18:1333-1344
37. Sahrman SA (1998) Diagnosis by the physical therapist—a prerequisite for treatment. A special communication. *Phys Ther* 68:1703-1706
38. Spitzer W (1987) Scientific approach to the assessment and management of activity-related spinal disorders. A monograph for clinicians. Report of the Quebec Task Force on Spinal Disorders. *Spine* 12:S1-S59
39. Strender LE, Sjoblom A, Sundell K, Ludwig R, Taube A (1997) Interexaminer reliability in physical examination of patients with low back pain. *Spine* 22:814-820

-
40. van den Hoogen HJ, Koes BW, Deville W, van Eijk JT, Bouter LM (1996) The inter-observer reproducibility of Lasegue's sign in patients with low back pain in general practice. *Br J Gen Pract* 46:727-730
41. Walter SD (1984) Measuring the reliability of clinical data: the case for using three observers. *Rev Epidemiol Sante Publique* 32:206-211
42. Wilson L, Hall H, McIntosh G, Melles T (1999) Intertester reliability of a low back pain classification system. *Spine* 24:248-254