J Rehabil Med 2002; 34: 176–182

Taylor&Francis healthsciences

CLINICAL TESTS ON IMPAIRMENT LEVEL RELATED TO LOW BACK PAIN: A STUDY OF TEST RELIABILITY

Eva Horneij,[1] Bertil Hemborg,[1] Birgitta Johnsson[2] and Charlotte Ekdahl[2]

From the [1]Hälsoinvest Ramlösa Clinic, Ramlösa, Sweden, and [2]Department of Physiotherapy, Lund University, Lund, Sweden

The objectives of the study were, in a working population, to standardize and evaluate a set of clinical tests on impairment level related to the low back with reference to intra- and inter-rater reliability. The study was undertaken in two steps. In step 1, 15 tests were examined for inter-rater reliability by three pairs of physiotherapists and for intra-rater reliability by one physiotherapist. Intra-rater reliability was acceptable ($\kappa > 0.40$) for 14 of the 15 tests. Inter-rater reliability was acceptable for 7 of the 15 tests. In step 2, the tests, indicating a non-acceptable inter-rater reliability ($\kappa \leq 0.40$) were further standardized and re-tested by two of the physiotherapists. This further standardization procedure resulted in an acceptable inter-rater reliability for all of these tests. Clinical tests of a working population should preferably be performed by the same rater. However, when tests are performed by different raters, it is suggested that test procedures should be regularly standardized, and in pain provocation tests, the magnitude of the applied pressure should be checked regularly and compared with co-raters, in order to improve inter-rater reliability.

Key words: physical examination, low back pain, reliability, non-patient, nursing aide.

J Rehabil Med 2002; 34: 176–182

Correspondence address: Eva Horneij, RPT, MSc, Hälsoinvest Ramlösakliniken, SE-256 57 Ramlösa, Sweden. E-mail: eva.horneij@swipnet.se

Submitted September 24, 2001; Accepted December 3, 2001

INTRODUCTION

The prevalence and incidence of low back pain among nursing aides and assistant nurses working in the home care service, are high (1, 2). It is thus important, at an early stage, to identify individuals at high risk of developing low back pain and to detect signs that may indicate a pathophysiological process (3). Physiological as well as psychological and social factors have been reported to interact with low back problems (4–6). Thus, the importance of a multivariate investigation of work-related complaints, including clinical findings, has been pointed out (7). If, however, clinical tests are supposed to be sensitive signs of early bad health, they have to be valid and reliable. This applies even if the exact pathology of the perceived low back pain may be hard to define (8). A degenerated disc, for example, may provoke pain in some subjects but not in others, and moreover, a degenerated disc is more often painful in younger than in older persons (5, 9). A prevalence estimate of 40% of low back pain emanating partly from the facet joints and confirmed by double blocks has been documented (6). Still the criterion validity of a clinical examination, intending to provoke pain from these joints has not yet been demonstrated (6, 10). The validity of clinical tests of the sacroiliac joint has been questioned (11, 12), even if it seems that the sacroiliac joint has a regulatory function for the stability of the lumbar spine (13). As validity is highly affected by random measurement errors consequently, there is a need of further studies on both the reliability and the validity of specific clinical tests.

Clinical tests may also be used for the evaluation of different interventions. When assessing the effects of an intervention programme, the responsiveness of the measurement tools is highly related to reliability. Studies on the reliability of clinical tests of the low back and the sacroiliac region, have usually been performed on patient populations (14–18).

Few studies have investigated the reliability of different clinical tests for the back region on a non-patient population (19, 20). The results of these studies are contradictory. In order to attain an acceptable inter-rater reliability a careful standardisation of the tests has been demanded (18, 19, 21). However, the impact of standardization has, to our knowledge, not yet been studied.

The objectives of the study were, in a working population, to standardize and evaluate a set of clinical tests on impairment level related to low back pain, employed extensively by physiotherapists and physicians, with reference to intra- and inter-rater reliability.

METHOD

The present study constitutes one part of a larger project, aiming at preventing or reducing disorders in the neck, shoulders and back among home-care personnel and was approved by the Ethical Committee of the Faculty of Medicine, University of Lund, Sweden. All participants gave their written consent before participation.

The study was undertaken in two steps. After the first evaluation of step 1, some of the tests with non-acceptable reliability, i.e. with a kappa value (κ) ≤ 0.40 (fair or poor) were further standardized, and retested for inter-rater reliability in step 2.

Subjects

In both steps, the subjects participating were nursing aides and assistant nurses working at least 20 hours/week in the home-care services. Subjects were included irrespective of ongoing musculoskeletal symptoms i.e. both healthy subjects and subjects with symptoms from the low back were included. Pain, aches or discomfort from the low back at any

 ISSN 1650–1977

Table I. *Number of subjects (n) being examined by the different pairs of raters (Step 1:A/A, A/B, A/C, and B/C. Step 2: A/B). Mean, standard deviation (SD), median and range for age and body mass index (BMI = kg/m^2)**

	A/A $n = 18$ Step 1	A/B $n = 14$ Step 1	A/C $n = 13$ Step 1	B/C $n = 17$ Step 1	A/B $n = 22$ Step 2
Age					
Mean (SD)	40.6 (10.9)	43.2 (8.4)	44.5 (7.1)	40.7 (10.2)	40.9 (9.0)
Median	41.0	46.5	46.0	42.0	41.5
Range	21–59	29–53	32–53	20–57	26–57
BMI					
Mean (SD)	25.2 (3.3)	22.9 (2.9)	24.6 (5.8)	22.2 (2.5)	25.1 (3.3)
Median	24.0	22.0	24.0	22.0	24.5
Range	21–32	19–28	19–42	19–28	20–32

* No significant difference among the groups was found, concerning age or BMI.

time during the preceding 12 months were reported by 53% of the subjects being examined for inter-rater reliability (step 1), by 46% of the subjects being examined for intra-rater reliability and by 50% of the subjects being examined in step 2. The corresponding proportions for incapacitating pain in the low back during the preceding 12 months were 13%, 15% and 14%, respectively. Women who were pregnant and those who were on sick leave were excluded.

Procedure

The subjects were examined in step 1 by three experienced physiotherapists: A, B and C (range of experience 18–25 years), all trained in orthopaedic manual therapy and two (A and C) with the Swedish degree in manual medicine. Intra-rater reliability was calculated for rater A. In step 2, the physical examinations were undertaken by physiotherapists A and B. In both steps the physiotherapists were not aware of the test results of their co-rater. Subjects were instructed not to mention previous or ongoing pain or discomfort experienced around the low back region but to indicate when certain tests caused pain. In both steps the clinical tests were performed in the same order for every subject.

Step 1. All nursing aides and assistant nurses, working in the home-care services in a village in the southern part of Sweden, were invited to participate. Sixty-two subjects (89%) accepted and 44 of these subjects were examined for inter-rater reliability and 18 for intra-rater reliability. The clinical examination was performed at the same hour on two consecutive days. The number of subjects being examined, age and body mass index (BMI) are presented in Table I. To eliminate systematic errors, the first and second examinations were assigned equally among the physiotherapists.

Fifteen clinical tests of musculoskeletal disorders, related to the back, were examined. Initially, in a pilot study, the tests were standardized on 17 female nursing aides and assistant nurses by the three physiotherapists in common. The techniques used in the clinical tests and the methods used to evaluate the clinical findings were discussed and then recorded in a test manual compiled by one of the physiotherapists (A). The subjects included in the pilot study were not included in step 1 of the study.

Step 2. Ten clinical tests were further standardized in order to improve their reliability. For example, the pressures of palpation were calibrated on a bathroom balance. In a pilot study, the standardization procedure was done in several steps. Firstly, the tests were evaluated by each physiotherapist *separately* and then revised by the two physiotherapists together. Secondly, the tests were standardized and evaluated by the physiotherapists in common. Thirdly, this procedure was repeated until a consensus on the performance and evaluation of the tests was reached. Fourthly, the instructions in the test manual were recorded by the two physiotherapists together.

Twenty-two female nursing aides and assistant nurses were consecutively selected from an ongoing investigation of the work environment. They were examined by two physiotherapists (A and B) on the same day. The time interval between the examinations was about 30 minutes. Age and BMI are presented in Table I. The order of the first and second examinations was randomly distributed between A and B.

The subjects in the pilot test were not included in the study of reliability.

Clinical tests

In both steps, clinical tests of the back, aiming at the evaluation of pain and muscle length and frequently used in physiotherapy practice, were chosen for the assessment of reliability. The results were presented dichotomously as normal/not normal except for the tests of the muscle length, which were presented in three categories as normal, tight, and excessive in length.

Step 1

The subject in the prone position

Springing test. The physiotherapist applied a bilateral postero-anterior force on the transverse processes of L5–T7. These tests of L5–L3 were added together, so that if one, two or three of the springing tests of L5, L4 or L3 were positive the test was presented as "not normal". The same procedure was applied for the segments of L2–T11 and for T10–T7. Each test movement for the thoracic spine should be at right angles to the targeted facet joints while the test movement for the lumbar spine was performed in a postero-anterior direction. While testing the L5–L3 segments, the pressure was held for at least 20 seconds in order to detect a possible "delayed stretched pain" (22). The test is intended to give information about stiffness and pain (23). In this study the test was dichotomously evaluated as pain/no pain.

Palpation directed towards the piriformis muscle and its insertion. The muscle belly was palpated at the crossing of the lines from the lateral crista of the pelvis to the ischial tuberosity and from the posterior superior iliac spine to the greater trochanter of the femur. The insertion of the piriformis was palpated medially to the top of the greater trochanter of the femur. The test was rated dichotomously as pain/no pain.

Palpation directed towards the ilio-lumbar ligament. This test was performed unilaterally with a ventral pressure in the space between the iliac crest and the transverse processes of L5 and L4. The test was rated positive if local and/or radiating pain was reported.

The subject in the supine position

Muscle length. All tests were performed passively according to Janda (23) and rated as 0 = normal length, 1 = restricted length and 2 = excessive length.

The assessment of the length of the hamstrings was performed with the knee straight and the opposite leg resting on the couch, and fixed horizontally. Provided that no signs of sciatica were present, the muscles were considered tight if the hip flexion was less than about 80°. An excessive length was registered if the flexion was more than about 90°. The iliopsoas muscle was tested with the subject lying on the lower edge of the examining couch with the buttock on the edge. The opposite leg was flexed so far that the lumbar lordosis was extinguished and was held by the physiotherapist in this position. The leg being tested was allowed to hang free. The muscle length was considered tight if the femur did not reach the horizontal plane and excessive if the femur was beyond the horizontal plane. The muscle length of the rectus femoris was tested in the same position as the iliopsoas muscle and considered tight if the knee

Table II. *Step 1. Intra-tester reliability of clinical tests of the back region (rater A). Percentage agreement, positive findings in each examination (test occasion 1/test occasion 2), and the kappa coefficient with the 95% confidence interval (CI). wκ = weighted kappa. Number of subjects being examined. The weighted kappa is only identified for one pair of raters and was thus not calculated*

	Percentage agreement ($n = 18$)	Positive findings	Kappa coefficient (95% CI)
Pain provocation tests			
Springing test L5-L3	78	9/9	0.56 (0.18; 0.94)
Springing test L2-T11	89	9/9	0.78 (0.49; 1.0)
Springing test T10-T7	89	6/4	0.73 (0.39; 1.0)
Palpation			
Piriformis muscle left	94	6/5	0.87 (0.62; 1.0)
Piriformis muscle right	83	6/5	0.61 (0.21; 1.0)
Piriformis insertion left	89	6/4	0.73 (0.39; 1.0)
Piriformis insertion right	89	6/4	0.73 (0.39; 1.0)
Iliolumbar ligament left	89	10/9	0.78 (0.50; 1.0)
Iliolumbar ligament right	83	6/6	0.64 (0.27; 1.0)
Muscle length			
Iliopsoas left	50	11/12	0.18 wκ (−0.25: 0.61)
Iliopsoas right	72	9/10	0.44 wκ (0.03; 0.86)
Rectus fem. left	72	11/12	0.70 wκ (0.40; 1.0)
Rectus fem. right	89	9/9	0.84 wκ (0.61; 1.0)
Hamstrings left	72	7/6	0.60 wκ (0.30; 0.91)
Hamstrings right	78	6/6	0.64 wκ (0.31; 0.97)

could not passively get into about 90° flexion and excessive if the knee was passively flexed more than about 90°, with the femur still in the horizontal plane. The muscle length could not be tested in the position stipulated by Janda on two persons due to lumbar pain, and on one person due to an earlier operation of the hip. Two persons were missed.

Step 2

Tests in step 1, with a non-acceptable reliability ($\kappa \leq 0.40$), were further standardized and evaluated by rater A and B. For the purpose of calibrating the pressure being applied by the two raters during palpation, a bathroom balance was used. A consensus on what force as well as what area of the palpation finger should be applied to the different structures was agreed upon. The test results were registered in the same way as in step 1.

The subject in the prone position

The springing test. The direction of the applied pressure, as described above, was again emphasized. The force being applied was agreed to be about 110 N. This force could be slightly moderated depending on the configuration of the subject, whether thickset or slender.

Palpation directed towards the piriformis muscle. The force required for the palpation of the piriformis muscle was agreed to be about 40 N. The palpation was performed across the fibres of the muscle belly.

Palpation directed towards the ilio-lumbar ligament. Two fingers were placed above the angle between the iliac crest and the transverse processes of L5 and L4 and a pressure was exerted by placing the other hand on top of the two fingers. The force was agreed to be 110 N and maintained for at least 20 seconds. As for the springing test, the pressure could be slightly differentiated according to whether the subject was thickset or slender.

The subject in the supine position

Length of the iliopsoas muscle. In step 1, the opposite hip was flexed until the lumbar lordosis was extinguished. As this position was difficult to maintain, in step 2 the hip was maximally flexed and kept in this position by the subject and the physiotherapist together. The edge of the couch was carefully adjusted to the level of the lower part of the sacrum. The muscle length was considered normal if the femur could easily be aligned with the horizontal plane (±10°). This position was checked using a plastic goniometer.

Length of the rectus femoris muscle. The test position was the same as for the iliopsoas muscle. In step 2, an excessive length was registered only when the knee was easily flexed more than 90° (±10°).

Statistics

The intra-rater reliability in step 1 and inter-rater reliability in step 2 were assessed by the observed frequency of exact agreements, the percentage agreement and by Cohen's kappa (24). When the outcome could be registered on an ordinal scale by more than two alternatives the weighted kappa was used (25).

For analysis of the inter-rater reliability in step 1, the results from the three pairs of raters were summarized and the generalization of unweighted kappa was used (26). The prevalence of positive findings, assessed by the three pairs of raters, was calculated through the mean of the positive findings from their first and second rating.

Kappa (κ) provides an indication of agreement beyond chance (25). The kappa coefficient has a maximum of 1.0. A value of zero indicates agreement no better than chance. Negative values show worse than chance agreement. According to Altman (27), the kappa value was interpreted as follows: κ: <0.20 = Poor, κ: 0.21–0.40 = Fair, κ: 0.41–0.60 = Moderate, κ: 0.61–0.80 = Good, κ: 0.81–1.00 = Very good.

The one-way ANOVA test was applied when groups of subjects were compared with respect to age and BMI. The Bonferroni method was used to correct for type I errors.

Analyses of reliability with more than two outcome possibilities and the generalised kappa were performed using the statistical packages StatXact (version 3) and Stata (version 6) for Windows, respectively. All other analyses were done with SPSS for Windows (version 8.0).

RESULTS

Step 1

Intra-rater reliability (Table II). Agreement was acceptable ($\kappa > 0.40$) for 14 of the 15 tests. Reliability was poor for the test of muscle length of the left iliopsoas and moderate to very good for the rest of the tests. The percentage agreement for the 14 tests, with an acceptable kappa value, varied between 72% and 94%.

Inter-rater reliability (Table III). The kappa value was calculated for a total of 15 tests. Reliability was acceptable ($\kappa > 40$) for 7 of the tests and non-acceptable ($\kappa \leq 0.40$) for the other 8 tests. The percentage agreement for tests with an acceptable kappa value varied between 65% and 88%, and for those with an unacceptable kappa value the variation was between 56% and 80%.

Step 2

Inter-rater reliability (Table IV). The kappa value was acceptable for all the tests ($\kappa > 0.40$). The percentage agreement varied between 73% and 95%.

DISCUSSION

The intra-rater reliability was acceptable for all tests except for the test of the length of the left iliopsoas muscle. However, the results show that a routine standardization of the clinical tests related to the low back in a working population, as performed in step 1, was not enough to gain an acceptable inter-rater reliability for 7 of 15 tests. A further and more careful standardization of the test procedure was needed to reach an acceptable reliability.

The standardization procedure differed between step 1 and step 2. For experienced physiotherapists, the performance and the evaluation of the clinical tests might have seemed self-evident, though not sufficient. In step 2, the standardization procedure promoted awareness of the differences between raters in both the performance as well as in the evaluation of the tests, and the reliability was improved.

When comparing measurement agreement between raters of categorical data, different statistical methods have been proposed such as, for example, different correlation coefficients or the χ^2 test (27, 28). These methods describe the association between two variables and not the agreement (24, 27, 28). The simplest way to study agreement is to calculate the percentage agreement or the absolute agreement between raters, which, however, do not account for agreement by chance. The ideal indexes of concordance should correct for agreement by chance (27, 28) and for this purpose the calculation of the kappa value has been recommended as the chance-corrected proportional agreement (24, 27, 28). However, as the kappa value depends on the prevalence of positive findings, it may be misleading to compare kappa values from different studies (27). The ideal situation when using the kappa value for agreement, is a 50% prevalence of positive findings (30). As this is not always the case, it is important that the agreement between raters is interpreted in its context and not only by its kappa value (27, 31).

The precision of the kappa values is presented by their 95% confidence intervals. Despite an acceptable point estimate of the kappa value of, for example, the springing test L3–L5 ($\kappa = 0.41$) in step 1, the precision of the value is weak (CI: 0.12; 0.70). The 95% confidence interval was also calculated for intra-rater reliability, step 1 and inter-rater reliability, step 2. However, due to the small sample size, these intervals are not comparable with the intervals of the inter-rater reliability, step 1 (27, 29).

The kappa value of the muscle length of the iliopsoas and the right rectus femoris in step 2 was acceptable. However, the raters did not agree on six subjects (27%) which, in our opinion, is not a sufficient concordance. The prevalence of positive findings was higher for the right rectus femoris than for the right iliopsoas muscle, explaining the higher kappa value for the test of the rectus femoris muscle. Saur et al. (19) found a poor reliability for these tests when performed on different days. When, however, the tests were performed directly after one another, the inter-rater reliability was slightly improved, mainly

Table III. *Step 1. Inter-tester reliability of clinical tests of the back region. Pooled results for all pair of raters (A, B, and C). Percentage agreement, prevalence of positive findings, the generalised kappa coefficient and the 95% confidence interval (CI) of the kappa value. n = number of subjects being examined*

	n	Percentage agreement	Prevalence (%)	Kappa coefficient (95% CI)
Pain provocation tests				
Springing test L5–L3	44	73	36	0.41 (0.12; 0.70)
Springing test L2–T11	44	68	50	0.36 (0.07; 0.66)
Springing test T10–T7	44	61	30	0.12 (−0.18; 0.41)
Palpation				
Piriformis muscle left	44	80	17	0.28 (−0.02; 0.57)
Piriformis muscle right	44	75	31	0.41 (0.12; 0.70)
Piriformis insertion left	44	86	18	0.54 (0.25; 0.84)
Piriformis insertion right	44	84	31	0.63 (0.33; 0.92)
Iliolumbar ligament left	44	73	25	0.35 (0.06; 0.57)
Iliolumbar ligament right	44	80	22	0.31 (0.09; 0.60)
Muscle length				
Iliopsoas left	41	63	32	0.21 (−0.06; 0.48)
Iliopsoas right	40	65	34	0.43 (0.16; 0.71)
Rectus fem. left	41	56	29	0.03 (−0.22; 0.29)
Rectus fem. right	39	72	27	0.30 (0.00; 0.59)
Hamstrings left	41	88	23	0.68 (0.45; 0.92)
Hamstrings right	41	88	23	0.68 (0.45; 0.92)

Table IV. *Step 2. Inter-tester reliability of clinical tests of the back region. Percentage agreement in number of patients with positive findings by each rater (rater A/rater B), and the kappa coefficient with the 95% confidence interval (CI). wκ = weighted kappa.* n = *number of subjects being examined. The weighted kappa is only identified for one pair of raters and was thus not calculated*

	Percentage agreement $n = 22$	Positive findings	Kappa coefficient (95% CI)
Pain provocation tests			
Springing test L2–T11	77	9/4	0.49 (0.15; 0.83)
Springing test T10–T7	77	8/8	0.49 (0.15; 0.83)
Palpation			
Piriformis muscle left	91	6/6	0.77 (0.47; 1.0)
Piriformis muscle right	82	6/4	0.49 (0.07; 0.91)
Iliolumbar ligament left	95	6/5	0.88 (0.65; 1.0)
Iliolumbar ligament right	77	7/8	0.50 (0.12; 0.88)
Muscle length			
Iliopsoas left	73	8/8	0.63 wκ (0.13; 0.84)
Iliopsoas right	73	8/6	0.56 wκ (0.07; 0.80)
Rectus fem. left	95	9/10	0.94 wκ (0.77; 1.0)
Rectus fem. right	73	10/10	0.70 wκ (0.24; 0.85)

for the iliopsoas muscles. It is possible that the muscle length varies naturally from one day to another, which was shown in the study by Hyytiäinen et al. (20), and that the improved reliability of the muscle length in step 2 was thus partly due to a shorter time interval between the different ratings.

In our opinion, the test position when testing the muscle length of the iliopsoas and rectus femoris muscles according to Janda (23), takes time to arrange and is difficult to standardize. Three persons in step 1 could not hold the proposed position due to pain or stiffness of the hip. From the point of view of reliability, the comfort of the test person and the time spent on the performance of the test, the tests of the muscle length of the iliopsoas and the rectus femoris would probably be favoured by a position different from that proposed by Janda. Moreover, the interpretation of the muscle lengths of the iliopsoas and rectus femoris in this study was classified, on the basis of common practice, as excessive, normal and restricted range. Arguments may be found against these strict limits, especially as 45% of the subjects measured in step 2 had a restricted muscle length of the rectus femoris. Saur et al. (19) assessed the inter-rater reliability of the muscle length of the iliopsoas and rectus femoris in the same positions as in the present study. Reliability was calculated for categories of degrees as well as continuously. They found reliability to be slightly better when muscle length was measured continuously. Thus, our study may have profited from the use of an easily applied inclinometer capable of continuous measurements (in degrees) of the muscle length.

Most of the tests performed in this study were tests for pain provoked by palpation. Levin et al. (21) found a wide variation in the pressure applied by different raters when testing the sacroiliac joint on the same person, with a coefficient of variance of about 25%. They also found variations within raters in the pressure applied on different test occasions. The raters were, however, capable of maintaining a relatively constant pressure for 20 seconds. In our study, step 2, the magnitude of the pressure applied in each test was previously decided. Keating et al. (32) showed that physiotherapists trained in an awareness of the magnitude of the pressure being applied, were able to replicate the same force up to 1 month after training. This result was most evident in low forces, up to about 100 N.

To promote inter- and intra-rater reliability, the importance of standardizing the pressure applied, as suggested by Levin et al. (21) and as performed in our study, is thus important.

Generally, the short time interval between ratings in step 2 might have had an impact on the improved reliability, with subjects being biased from their first report. However, this impact must be greater if only a few tests are performed. In this study, apart from the 10 tests of the back in step 2, 24 tests of the neck and shoulders were performed (unpublished data) and thus the possibility of recall bias should be small.

In the study by Strender et al. (18), on a patient population, a non-acceptable, inter-rater reliability of the springing test was found, despite a standardized test procedure having been performed, within a 30-minute interval, by experienced physiotherapists. In our study of the springing tests in step 2, the kappa value was acceptable.

In a study on healthy subjects, Hogeweg et al. (33) showed a great variation in pain threshold, continuously measured, in palpations performed 30 mm laterally of the spinous processus when the tests were repeated directly after one another. Concordance between raters was, however, found acceptable when the means of three repeated measurements were compared. In the present study, the number of repeated tests was not standardized. More than one single test might thus have been performed by the rater before the final test result was recorded in the test manual, i.e. one rater might have registered her first test while the other rater registered her second trial. As categorical data were used, it is possible that agreement of the pain provocation tests would have improved if consensus were reached on which test trial is to be registered, as for example, always the first test.

Kosek et al. (34) found that women with no musculoskeletal

problems, reported different pressure pain thresholds when the tests were either immediately repeated or repeated after 20–30 minutes. Thus even the patient's interpretation of the pain pressure on different occasions might be a source of measurement error.

In the early prevention of low back pain, especially in connection with physically demanding work, disorders may not be manifest. Thus, in early prevention, a clinical examination will mostly rely on the quantification of single clinical tests.

It is obvious that clinical tests of the low back performed in a working, non-patient population, need to be carefully standardized in order to reach an acceptable reliability. In this study, however, despite an acceptable kappa value for all tests in step 2, the raters disagreed in 5 of the subjects (23%) for 2 of the pain palpation tests and in 6 of the subjects (27%) for the tests of the muscle length of the iliopsoas and rectus femoris (Table IV). Thus, the value of performing only single tests in order to identify individuals at high risk of developing low back pain may be questioned. However, in early prevention, the results of several clinical tests might form a pathophysiological pattern that could improve agreement, which should be further studied, as should the efficacy of diagnostic marker tests, i.e. the sensitivity and the specificity of these tests (28).

CONCLUSION

In a working population of nursing aides and assistant nurses, the inter-rater reliability of clinical tests related to the back, was acceptable ($\kappa > 0.40$) for only 7 out of 15 tests, despite a routine standardization procedure. A further and more careful standardization procedure of the 8 tests with a non-acceptable, inter-rater reliability resulted in an acceptable reliability for all these tests. However, despite an acceptable kappa value of the muscle length of the iliopsoas and rectus femoris, the raters did not agree on 6 subjects (27%), which is not sufficient concordance. Clinical tests of a non-patient population should preferably be performed by the same rater. However, when tests are performed by different raters, it is suggested that test procedures should be regularly standardized and in pain provocation tests, the magnitude of the applied pressure should be checked regularly and compared with co-raters, in order to improve inter-rater reliability.

ACKNOWLEDGEMENTS

Gratitude is expressed to the AMF-trygghetsförsäkring for financial support which made this research possible. We also thank the physiotherapists Lena Perhag and Birgitta Österberg and the home-care personnel who kindly consented to participate in this study.

REFERENCES

1. Ono Y, Lagerström M, Hagberg M, Lindén A, Malker B. Reports of work-related musculoskeletal injury among home care service workers compared with nursery school workers and the general population of employed women in Sweden. Occup Environ Med 1995; 52: 686–693.
2. Pheasant S, Stubbs D. Back pain in nurses: epidemiology and risk assessment. Appl Ergonomics 1992; 23: 226–232.
3. Delitto A. Are measures of function and disability important in low back care? Phys Ther 1994; 74: 452–462.
4. Bongers P, de Winter C, Kompier M, Hildebrandt V. Psychosocial factors at work and musculoskeletal disease. Scand J Work Environ Health 1993; 19: 297–312.
5. Vanharanta H, Guyer R, Ohnmeiss D, Stith W, Sachs B, Aprill C, et al. Disc deterioration in low-back syndromes. A prospective, multicenter CT/discography study. Spine 1988; 13: 1349–1351.
6. Schwarzer C, Wang S, Bogduk N, McNaught P, Laurent R. Prevalence and clinical features of lumbar zygapophysial joint pain: a study in an Australian population with chronic low back pain. Ann Rheum Dis 1995; 54: 100–106.
7. Elert J, Brulin C, Gerdle B, Johansson H. Mechanical performance, level of continuous contraction and muscle pain symptoms in home care personnel. Scand J Rehabil Med 1992; 24: 141–150.
8. Vällfors B. Acute, subacute and chronic low back pain: clinical symptoms, absenteeism and working environment. Scand J Rehabil Med 1985; Suppl 11: 27–58.
9. Vanharanta H, Sachs B, Ohnmeiss D, Aprill C, Spivey M, Guyer R, et al. Pain provocation and disc deterioration by age. A CT/discography study in a low-back pain population. Spine 1989; 14: 420–423.
10. Schwarzer A, Derby R, Aprill C, Fortin J, Kine G, Bogduk N. Pain from the lumbar zygapophysial joints: a test of two models. J Spinal Disord 1994; 7: 331–336.
11. Dreyfuss P, Michaelsen M, Pauza K, McLarty J, Bogduk N. The value of medical history and physical examination in diagnosing sacroiliac joint pain. Spine 1996; 21: 2594–2602.
12. Maigne JY, Aivaliklis A, Pfefer F. Results of sacroiliac joint double block and value of sacroiliac pain provocation tests in 54 patients with low back pain. Spine 1996; 21: 1889–1892.
13. Indahl A, Kaigle A, Reikerâs O, Holm S. Sacroiliac joint involvement in activation of the porcine spinal and gluteal musculature. J Spinal Disord 1999; 12: 325–330.
14. Waddell G, Main C, Morris E, Venner R., Rae P, Sharmy S, et al. Normality and reliability in the clinical assessment of backache. Br Med J 1982; 284: 1519–1523.
15. McCombe F, Fairbank J, Cockersole B, Pynsent P. Reproducibility of physical signs in low-back pain. Spine 1989; 14: 908–918.
16. Laslett M, Williams M. The reliability of selected pain provocation tests for sacroiliac joint pathology. Spine 1994; 19: 1243–1249.
17. Maher C, Adams R. Reliability of pain and stiffness assessment in clinical manual lumbar spine examination. Phys Ther 1994; 74: 801–809.
18. Strender LE, Sjöblom A, Sundell K, Ludwig R, Taube A. Inter-examiner reliability in physical examination of patients with low back pain. Spine 1997; 22: 814–820.
19. Saur P, Pfingsten M, Ensink FB, Heinemann R, Koch D, Seeger D, et al. Interrater-Untersuchungen zur Reliabilitätsprüfung somatischer Befunde. Rehabilitation Stuttg 1996; 35: 150–160.
20. Hyytiäinen K, Salminen J, Suvitie T, Wickström G, Pentti J. Reproducibility of nine tests to measure spinal mobility and trunk muscle strength. Scand J Rehabil Med 1991; 23: 3–10.
21. Levin U, Nilsson-Wikmar L, Stenström C, Lundeberg T. Reproducibility of manual pressure force on provocation of the sacroiliac joint. Physiother Res Int 1998; 3: 1–14.
22. Harms-Ringdahl K, Brodin H, Eklund L, Borg G. Discomfort and pain from loaded passive joint structures. Scand J Rehabil Med 1983; 15: 205–211.
23. Lewit K. Manipulative therapy in rehabilitation of the locomotor system. Oxford: Butterworth-Heinemann; 1991.
24. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37–46.
25. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70: 213–220.
26. Fleiss J. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76: 378–382.
27. Altman D. Practical statistics for medical research. 4th ed. London: Chapman & Hall; 1994. p. 403–409.

28. Kramer M, Feinstein A. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther 1981; 29: 111–123.
29. Fleiss J, Cicchetti D. Inference about weighted kappa in the non-null case. Appl Psychol Meas 1978; 2: 113–117.
30. Lantz C. Application and evaluation of the kappa statistic in the design and interpretation of chiropractic clinical research. J Manipulative Physiol Ther 1997; 20: 521–552.
31. Haas M. The reliability of the reliability. Review of the literature. J Manipulative Physiol Ther 1991; 14: 199–208.
32. Keating J, Matyas T, Bach T. The effect of training on physical therapists' ability to apply specified forces of palpation. Phys Ther 1993; 73: 45–53.
33. Hogeweg J, Langereis M, Bernards A, Faber J, Helders P. Algometry. Measuring pain threshold, method and characteristics in healthy subjects. Scand J Rehabil Med 1992; 24: 99–103.
34. Kosek E, Ekholm J, Nordemar R. A comparison of pressure pain thresholds in different tissues and body regions. Long-term reliability of pressure algometry in healthy volunteers. Scand J Rehabil Med 1993; 25: 117–124.