Intra-observer and inter-observer agreement of the manual examination of the lumbar spine in chronic low-back pain

Etienne Qvistgaard · Jens Rasmussen ·
Jes Lætgaard · Steen Hecksher-Sørensen ·
Henning Bliddal

Received: 20 March 2006 / Accepted: 15 April 2006 / Published online: 6 May 2006
© Springer-Verlag 2006

Abstract Examination is a cornerstone in the manual procedures leading to mobilisation/manipulation of the low back. The observer variation of the more specific segmental tests remains to be investigated. Two skilled specialists in manual medicine examined the segmental changes in the lumbar spine. The patients were unknown to the examiners and no information of the case history was given. All test results were recorded by an observer present in the room who ensured that no conversation was allowed during the examination. The primary outcome measures were the kappa values for each test. The matching was defined as acceptable (acc) within two neighbouring levels and perfect (per) on the same level. Intra-observer variation (tested in 33 patients and 10 subjects without low-back pain): The agreement between first and second segmental diagnosis examination was 70% (per) and 82% (per + acc). Kappa values were: segmental diagnosis 0.60 (per) and 0.70 (per + acc), multifidus test 0.51 (per) and

0.60 (per + acc), sideflexion 0.57 (per) and 0.69 (per + acc), and ventral flexion 0.31 (per) and 0.45 (per + acc). Inter-observer variation (tested in 60 patients): The agreement for segmental diagnosis between the examiner A and B was 42% (per) and 75% (per + acc). Kappa values were: segmental diagnosis 0.21 (per) and 0.57 (acc), multifidus test 0.12 (per) and 0.48 (acc), sideflexion 0.22 (per) and 0.45 (acc), and ventral flexion 0.22 (per) and 0.44 (acc). By manual tests, skilled examiners seem to be able to diagnose segmental dysfunctions in the low back. The clinical implication of these dysfunctions remains to be clarified.

Keywords Low-back pain · Clinical examination · Observer variation · Kappa · Segmental dysfunction

Introduction

In most cases, chronic low-back pain (LBP) is regarded as non-specific [14]. Even so, it is generally agreed that it is very important for the course of the disease that the patient feels reassured by a thorough examination [10]. However, the clinical examination is difficult with a very varying reproducibility of key elements of the objective evaluation of the patient [3].

More specialised tests for localised tension and reduced mobility have not been similarly successful [3, 4, 16] although the results may improve with specialisation of the examiners [17, 20]. Very few results have been presented to support the possibility of an exact clinical diagnosis of a more specific pathology in the low back [21]. To a certain degree, this contrasts the fact that segmental manipulation based on such

Planning of study and protocol: Henning Bliddal, Jens Rasmussen, and Etienne Qvistgaard, Inclusion of patients: Henning Bliddal and Jens Rasmussen, Manual examination: Jes Lætgaard and Steen Hecksher-Sørensen, Statistical evaluation: Etienne Qvistgaard and Henning Bliddal, Preparation of manuscript: all authors

E. Qvistgaard · J. Rasmussen · H. Bliddal (✉)
The Parker Institute, H:S Frederiksberg Hospital,
2000 Frederiksberg, Denmark
e-mail: henning.bliddal@fh.hosp.dk

J. Lætgaard · S. Hecksher-Sørensen
Private Clinic of Manual Medicine, Horsens,
8700 Horsens, Denmark

examination is a well accepted and effective therapy of LBP [8, 23].

The aim of this study was to test the observer variation of the manual clinical examination of the lumbar spine with regard to specific segmental tests.

Materials and methods

Subjects

The general practitioners in the district were asked to refer patients with chronic LBP to the LBP out-patients' clinic at the Department of Rheumatology, Frederiksberg Hospital.

The criteria of inclusion were: age between 18 and 60 and LBP for more than 1 month. The criteria of exclusion were: clinical signs of an acute disc herniation, inflammatory disease including Bechterew's disease, ongoing insurance claim, significant medical disease, intellectual or language problems. The patients were consecutively enrolled in the two parts of the study. In the intra-observer study, volunteers from the Department participated as healthy controls.

Methods

The manual examination focused mainly on segmental hypomobility and abnormal muscle tension (summarised in the phrase '*dysfunction*'), while verbal pain reactions were not tested for due to the blinding of the investigator. The investigators were only allowed to describe a given test as normal or abnormal (binary data). All subjects were tested for the same signs of lumbar involvement [7]:

Standing: trunk side bending, flexion, 'stork test'

Supine: pelvic girdle (clockwise and counter clockwise rotation)

Prone: lumbar spring-test with palpation of the interspinous movement and multifidus test (Fig. 1)

Positioned on the side with the front to the examiner: extension, flexion, and rotation tested by palpation of movement of neighbouring segments (Fig. 1).

The tests included the segments Th12–L1, L1–L2, L2–L3, L3–L4, L4–L5, and L5–S1. In the intra-observer part of the study, the segmental level was defined by the examiner separately at each test. In the inter-observer part of the study, the processus spinosi of L1 and L5

Fig. 1 Manual tests for movement of neighbouring processus spinosi by **a** extension test (springing test), **b** multifidus test, and **c** rotation. The hypomobility observed at the examination was given a final diagnosis and **d** a segmental manipulation performed

were marked with red ink by the first examiner to diminish the variation due to different interpretation of the spinous level.

After the tests, the investigator was asked to summarise his observations into one final diagnosis of the most pronounced segmental disturbance.

Intra-observer reliability

For the study of intra-tester variation, healthy controls from the staff unknown to the tester were recruited. A total of 33 patients and 10 controls (10 male and 23 female participants, mean age 45 years, range 21–61 years) were tested in groups of 3–4 participants at a time. The investigator (J.L.) was blindfolded (Fig. 2) and tested all subjects twice in random succession. Another investigator took down notes of the examinations and made certain that the blindfolding was stable during the tests. No conversation was allowed during the examination and all instructions were given by the assistant. The tests only included palpation: flexion/extension/rotation with the subject positioned on the side and prone, and multifidus test.

Inter-observer reliability

A total of 60 patients were tested in separate rooms by two investigators in random succession.

During the examinations, no conversation was allowed except instructions of positioning and manoeuvres of the patient. An investigator who did not participate in the examination was present in the room during the procedure and took down notes of the results.

The patients were all unknown to the investigators performing the examinations.

Fig. 2 Blindfolded examiner performing a flexion test in a subject. The examiner was not allowed any conversation with the patient who was instructed by the independent observer

Statistics

Perfect agreement (per) was defined as positive result on the same segmental level whereas acceptable agreement (acc) was defined as positive result on neighbouring segmental levels. Kappa statistics were performed including the results of each of the six segments (Fig. 3). The kappa values were evaluated as reasonable over 0.4, good over 0.6, and excellent over 0.8.

Ethical considerations

The local ethical committee approved the study.

Results

Intra-observer variation

The agreement between first and second examinations was 70% (per) and 82% (per + acc). Kappa values by the two definitions of matching are shown in Table 1. All kappa values were in the acceptable range. The most reproducible was the final diagnosis while ventral flexion (patient lying on the side) had the least reproducible results.

All patients but one had dysfunctions at certain levels; however, also, eight of ten of the healthy control subjects with no history of sick leave for LBP had a dysfunction diagnosed.

Inter-observer variation

The agreement between the examiner A and B was 42% (per) and 75% (per + acc). The kappa values are given (shown) in Table 2. A skewness of the diagnostic pattern was noted between the two examiners with a difference in the most often observed level of diagnosis (Fig. 4).

Discussion

In the present study, skilled examiners were able to identify the same segmental level of spinal dysfunction in the lumbar spine by manual investigation in the main part of the patients with chronic LBP. The level was very reproducible in the intra-observer study and was within acceptable ranges for inter-observer variation. Some differences might have been due to a discrepancy in the reporting of the level of testing. This was indicated by the many diagnoses on neighbouring

Fig. 3 In order to dichotomise the results for kappa processing, two models were used. Perfect agreement only includes diagnoses of the exact same segmental level as chosen by the two investigators. Acceptable agreement allows displacement of the depicted segmental level by the level above or below. This is a constructed example of the model: **A** duplicate line indicates the segment chosen by the investigator 1. **a** In case of Th12/L1 or

L5/S1, there are only two possible levels to be chosen by the investigator 2 as 'acceptable' agreement including the perfect one. **b** In case of investigator 1 choosing one of the mid-four levels L1/2 to L4/5, in this case L3/4, there are three possible levels left to be included as 'acceptable' agreement including the perfect one

Table 1 Kappa values for intra-observer examination of the low back ($n = 43$)

Test	Kappa perfect match	Kappa acceptable match
Mobilisation in extension	0.54	0.64
Sideflexion	0.57	0.69
Ventral flexion	0.31	0.45
Multifidus test	0.51	0.60
Diagnosis (final)	0.60	0.70

Table 2 Kappa values for inter-observer examination of the low back ($n = 60$)

Test	Kappa perfect match	Kappa acceptable match
Mobilisation in extension	0.23	0.52
Sideflexion	0.22	0.45
Ventral flexion	0.22	0.44
Multifidus test	0.12	0.48
Diagnosis (final)	0.21	0.57

segments in the inter-observer part of the study, here given as acceptable values for segmental diagnoses. In several previous studies, observer variation in low-back

Fig. 4 The final diagnosis of segmental changes by examiners A and B. The diagnosis had to be chosen at one of the levels between Th12/L1 and L5/S1

examination has been tested with very varying kappa values according to the techniques applied. Values as high as 0.80 or more have been associated with more general diagnoses of dysfunction [19, 21], while low to moderate results in the range below 0.50 and even negative values have been reported of the more specific segmental examination [5, 18]. The present study

put the manual skills to a further and more elaborate test employing investigators who had been working together for years, using parallel techniques. The investigators were not allowed any information from the patient about the nature of the back problem. Thus, the diagnosis of a certain segmental dysfunction as well as the final segmental diagnosis was based on the manual examination only. In the intra-observer part of the study, the investigator found segmental dysfunctions in almost all participants without differences between patients and control subjects. According to the investigator, the dysfunctions in some of the controls were quite pronounced corresponding to severe mobility problems at this level of the back. This may be interpreted in several ways. Some of the controls might not be healthy persons, but rather ‘non-patients’, who either disregard the troubles of their back, or for some reason, do not react with pain to this back-disorder. The observation of dysfunctions in practically all subjects in this age group is in agreement with the frequent signs of degeneration in the spine [1].

It also stresses the importance of a careful analysis of the case history in comparison with objective signs of back pathology before any intervention is offered to a patient. With the possible exception of herniated discs, very few cases of chronic LBP may be fully clarified by imaging techniques. As a consequence, therapeutic measures are mostly recommended as generalised intervention as e.g. training exercises or behavioural back school [12, 22]. Manipulation directed against specific dysfunctions in the low back is widely used and reported by the patients to be of a definite value in the therapy of LBP [15].

A prerequisite for specific therapy of segmental dysfunction in the spine would be a reproducible diagnosis of the segmental level against which to direct the therapy. This could not be concluded from a previous test of precise diagnosis of segmental dysfunctions because of low to moderate kappa values in the range 0.27–0.47 [5]. Although a significant inter-observer association was demonstrated, the two examiners had somewhat different approaches to the examination as exemplified by the systematic variation in the levels of diagnosis, which is also experienced in former studies [5, 8]. This is reflected in the clinical experience of different treatment strategies, and as long as no ‘bottom line’ method for examination has been found, no general guidelines for manual therapy can be given.

The interest in the multifidus muscles is increasing with several studies suggesting a specific role of this muscle in the pathogenesis of back pain [9, 11]. In the present study, a simple manual estimate of muscle

tension gave remarkably good kappa value for the intra-observer test with a reasonably acceptable value for the inter-observer kappa. This test points towards a specific contraction of this muscle in the affected area. The contraction may be secondary to a number of painful changes occurring in the spine and cannot be taken into account for any specific disease mechanism. However, an objective determination of the specific level of pain and muscle dysfunction might increase the specificity of studies of multifidus, which have hitherto reported moderate changes in the average size of the multifidus muscle at only one level in patients with LBP [2, 9].

The present study was conducted to test if a consensus of diagnosis could be reached by palpatory examination only. The positive results are in some contrast to previous studies of the clinical examination of, e.g. sacroiliac joints, which have come to the conclusion that pain provoking tests are superior to other kinds of examination [13]. Studies of the examination of the lumbar spine have reported varying results [6, 21]. In most cases, a high degree of inter-observer variation has questioned the validity of the tests. In our study, the intra- and inter-observer variations were in the same range. It should be realised that the present results were obtained with observers with more than 20 years of experience in the examination of the low back and with years of coordination between them in a specialised clinic.

The manual examination of the lumbar spine may add new aspects to the study of diagnosis and intervention. However, it is still to be clarified which structures are involved in a dysfunction. Although a multifidus tension in this segment might indicate a pain reaction, a meaningfulness of the findings is weakened by the fact that such tension also exists in subjects without LBP. It will require further studies using the specific segmental diagnosis preferably in patients with shorter lasting LBP to demonstrate any importance per se of a specific diagnosis and of therapies directed against these dysfunctions.

Acknowledgments This study was supported by the Oak Foundation and The Danish Health Foundation.

References

1. Biering-Sorensen F, Hansen FR, Schroll M, Runeborg O (1985) The relation of spinal X-ray to low-back-pain and physical-activity among 60-year-old men and women. *Spine* 10:445–451
2. Danneels LA, Vanderstraeten GG, Cambier DC, Witvrouw EE, De Cuyper HJ (2000) CT imaging of trunk muscles in chronic low back pain patients and healthy control subjects. *Eur Spine J* 9:266–272

3. Donahue MS, Riddle DL, Sullivan MS (1996) Intertester reliability of a modified version of McKenzie's lateral shift assessments obtained on patients with low back pain. *Phys Ther* 76:706–716
4. Dreyfuss P, Michaelsen M, Pauza K, McLarty J, Bogduk N (1996) The value of medical history and physical examination in diagnosing sacroiliac joint pain. *Spine* 21:2594–2602
5. French SD, Green S, Forbes A (2000) Reliability of chiropractic methods commonly used to detect manipulable lesions in patients with chronic low-back pain. *J Manipulative Physiol Ther* 23:231–238
6. Gonnella C, Paris SV, Kutner M (1982) Reliability in evaluating passive intervertebral motion. *Phys Ther* 62:436–444
7. Greenman PE (1996) Principles of manual medicine. Williams & Wilkins, Baltimore, pp 99–103
8. Hawk C, Phongphua C, Bleecker J, Swank L, Lopez D, Rubley T (1999) Preliminary study of the reliability of assessment procedures for indications for chiropractic adjustments of the lumbar spine. *J Manipulative Physiol Ther* 22:382–389
9. Hides JA, Stokes MJ, Saide M, Jull GA, Cooper DH (1994) Evidence of lumbar multifidus muscle wasting ipsilateral to symptoms in patients with acute/subacute low back pain. *Spine* 19:165–172
10. Indahl A, Velund L, Reikeraas O (1995) Good prognosis for low back pain when left untampered. A randomized clinical trial. *Spine* 20:473–477
11. Indahl A, Kaigle AM, Reikeras O, Holm SH (1997) Interaction between the porcine lumbar intervertebral disc, zygapophysial joints, and paraspinal muscles. *Spine* 22:2834–2840
12. Lankhorst GJ, Vandestadt RJ, Vogelaar TW, Vanderkorst JK, Prevo AJH (1983) The effect of the Swedish Back School in chronic idiopathic low-back-pain—a prospective controlled-study. *Scand J Rehabil Med* 15:141–145
13. Laslett M, Williams M (1994) The reliability of selected pain provocation tests for sacroiliac joint pathology. *Spine* 19:1243–1249
14. Leboeuf-Yde C, Lauritsen JM, Lauritzen T (1997) Why has the search for causes of low back pain largely been non-conclusive? *Spine* 22:877–881
15. Meade TW, Dyer S, Browne W, Townsend J, Frank AO (1990) Low-back-pain of mechanical origin—randomized comparison of chiropractic and hospital outpatient treatment. *Br Med J* 300:1431–1437
16. Meijne W, van Neerbos K, Aufdemkampe G, van der WP (1999) Intraexaminer and interexaminer reliability of the Gillet test. *J Manipulative Physiol Ther* 22:4–9
17. Nelson MA, Allen P, Clamp SE, de Dombal FT (1979) Reliability and reproducibility of clinical findings in low-back pain. *Spine* 4:97–101
18. Phillips DR, Twomey LT (2000) A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. *Man Ther* 1:82–87
19. Razmjou H, Kramer JF, Yamada R (2000) Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low-back pain. *J Orthop Sports Phys Ther* 30:368–383
20. Riddle DL, Rothstein JM (1993) Intertester reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine* 18:1333–1344
21. 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
22. van Tulder M, Malmivaara A, Esmail R, Koes B (2000) Exercise therapy for low back pain—a systematic review within the framework of the Cochrane Collaboration Back Review Group. *Spine* 25:2784–2796
23. Waddell G (1995) Modern management of spinal disorders. *J Manipulative Physiol Ther* 18:590–596