Journal of Back and Musculoskeletal Rehabilitation 30 (2017) 551–557
DOI 10.3233/BMR-150495
IOS Press

Inter-rater reliability of diagnostic criteria for sacroiliac joint-, disc- and facet joint pain

Cornelis W.J. van Tilburg[a,*], Johannes G. Groeneweg[b], Dirk L. Stronks[b] and Frank J.P.M. Huygen[b]
[a]*Multidisciplinary Pain Center, Department of Anesthesiology, Bravis Hospital, Bergen op Zoom, The Netherlands*
[b]*Center for Pain Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands*

Abstract.
BACKGROUND/OBJECTIVE: Several diagnostic criteria sets are described in the literature to identify low back pain subtypes, but very little is known about the inter-rater reliability of these criteria. We conducted a study to determine the reliability of diagnostic tests that point towards SI joint-, disc- or facet joint pain.
METHODS: Inter-rater reliability study alongside three randomized clinical trials. Multidisciplinary pain center of general hospital. Patients aged 18 or more with medical history and physical examination suggestive of sacroiliac joint-, disc- and facet joint pain on lumbar level. Making use of nowadays most common used diagnostic criteria, a physical examination is taken independently by three physicians (two pain physicians and one orthopedic surgeon). Inter-rater reliability (Kappa (κ) measure of agreement) and significance (p) between raters are presented. Strengths of agreement, indicated with κ values above 0,20, are presented in order of agreement.
RESULTS: One hundred patients were included. None of the parameters from the physical investigation had κ values of more than 0.21 (fair) in all pairs of raters. Between two raters (C and D), there was an almost perfect agreement on three parameters, more specifically "Abnormal sensory and motor examination, hyperactive or diminished reflexes", "Sitting exam shows no reflex, motor or sensory signs in the legs" and "Straight leg raising (Laségue) negative between 30 and 70 degrees of flexion". The "Drop test positive" parameters had moderate strength of agreement between raters A and D and fair strength between raters A and B. The "Digital interspinous pressure test positive" had moderate strength of agreement between raters C and D and fair strength of agreement between raters A and B as well as raters B and C. Three other parameters had a fair strength of agreement between two raters, all other parameters had a slight or poor strength of agreement. Inter-rater reliability, confidence intervals and significance of pooled items for SI joint-, disc- and facet joint pain are represented; κ values for the pooled parameters of the physical examination suggestive of SI joint pain stayed below 0.20 between all raters. The same applies for the pooled parameters of the physical examination suggestive of facet joint or disc pain.
CONCLUSIONS: The poor reliability of the diagnostic parameters seriously limits their predictive validity, and as such their use in patients with low back pain for more than 3 months.

Keywords: Reliability and validity, reliability of results, diagnostic equipment, low back pain, sacroiliac joint, facet joint

1. Background

The assessment and interpretation of tests used to diagnose low back pain subtypes are often not standardized; however, this is necessary for the testing to be both valid and reliable [1]. Until now little is known about the inter-rater reliability of these diagnostic criteria. Regarding the diagnostic criteria, Young et al. demonstrated that pain when rising from sitting, as well as centralization of pain was associated with discogenic pain and that absence of pain when rising from sitting was associated with facet joint pain; sacroiliac (SI) joint pain was associated with three or more positive pain provocation tests, pain when ris-

*Corresponding author: Cornelis W.J. van Tilburg, Consultant Anesthesiologist & Pain Specialist, Multidisciplinary Pain Center, Department of Anesthesiology, Bravis Hospital, Boerhaaveplein 1, 4624 VT Bergen op Zoom, The Netherlands. Tel.: +31 887067697, +31 887067698; Fax: +31 887067699; E-mail: vtilburg@ziggo.nl.

ISSN 1053-8127/17/$35.00 © 2017 – IOS Press and the authors. All rights reserved
This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License (CC BY-NC 4.0).

Table 1
Findings from the physical examination suggestive of a SI [1–8] –, disc [2,3] – or facet joint [2,3] pain

Physical examination	SI	Disc	Facet
Drop-test positive	√		
Sitting exam shows no reflex, motor or sensory signs in the legs	√		
Straight leg raising (Laségue) negative between 30 and 70 degrees of passive flexion	√		
Distraction (Gapping) test positive	√		
Posterior shear (thigh trust) test positive	√		
Pelvic torsion (Gaenslen's) test positive	√		
Patrick-Faber test positive	√		
Compression test positive	√		
Sacral thrust test positive	√		
Cranial shear test positive	√		
Bilateral internal rotation of the hip/Unilateral rotation of the hip painful at SI joint(s)	√		
Yeoman's test positive	√		
Gait deviation		√	
Abnormal sensory and motor examination, hyperactive or diminished reflexes		√	
Digital Interspinous Pressure (DIP) test positive		√	
Straight leg raising (Laségue) positive between 30 and 70 degrees of passive flexion		√	
Pain in extension			√
Pain eased in flexion			√
Pain when rising from forward flexion			√
Schober test $<$ 3–5 cm			√
Pain in extension, lateral flexion or rotation manoeuvers to the ipsilateral side			√
Replication or aggrevation of pain by unilateral pressure over the ipsilateral side			√
Local unilateral passive movements show reduced range of motion or increased stiffness on the side of the involved facet joints			√
Tight or facilitated muscles (psoas, hip adductors, gluteus medius muscles)			√
Weak muscles (gluteus maximus, gluteus medius)			√

ing from sitting, unilateral pain and absence of lumbar pain [2]. In a systematic review to determine the diagnostic accuracy of tests available to clinicians to identify the source of low back pain, Hancock et al. found that centralization was the only clinical feature to increase the likelihood of the disc as being the source of pain, while absence of degeneration on MRI decreased this likelihood. A combination of SI joint tests was informative, single tests not [3].

We conducted this study to determine the reliability of diagnostic tests that point towards SI joint-, disc- or facet joint pain. The diagnostic tests mentioned in the literature on this subject were used.

2. Methods

We conducted an inter-rater reliability study in patients aged 18 years or more with low back pain for more than 3 months, who were referred to the pain center of a general hospital. The guidelines for reporting of studies of reliability and agreement (GRRAS [4]) were followed.

Patients with a suspicion of having a spine related pain disorder on lumbar level who met the inclusion – (age $\geqslant$ 18 years, chronic ($>$ 3 months) low back pain) and exclusion (presence of red flags, progressive neurological deficits, major psychiatric disorder (according to psychiatrists opinion), pain in other parts of the body that is more severe, pregnancy, active infection, communication (language) difficulties (according to physicians opinion)) criteria were eligible for inclusion. A total of three pain physicians and one orthopedic surgeon participated in the trial. The examination for each individual patient was performed by a combination of two pain physicians and one orthopedic surgeon. The consultations took place within a period of two weeks to decrease the chance for confounding and jointly determine the cause of the pain problem. A training session was held before the study to ensure as much consistency as possible of methods and standardization of test procedures, during which every item from the list with diagnostic criteria were judged on their presence or absence (Table 1). Before the physical examinations took place medical history was noted. The diagnostic criteria as well as the raters were applied in randomized order. The first pain physician that questioned and examined the patient also took into account the results from spinal imaging. Each physician made a working diagnosis in each patient. If the working diagnoses from the three physicians were in agreement with each other, a general working diagnosis was made, after which a diagnostic test block was performed. The study flowchart is presented in Fig. 1.

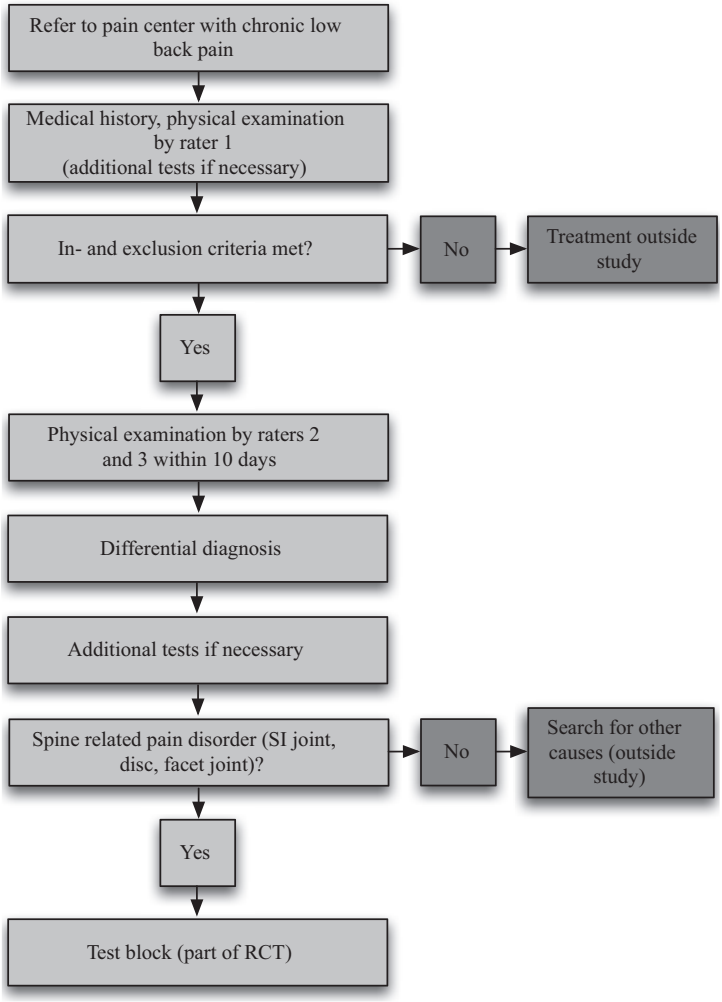

Fig. 1. Study flowchart.

The medical ethics committee from Erasmus University Medical Center approved the protocol. Written informed consent was obtained from all patients.

Data were analyzed using SPSS for Windows, version 22 (International Business Machines (IBM) Corporation, Software Group, Route 100, Somers, NY, 10589, United States of America). Inter-rater reliability of nowadays most common used diagnostic criteria was estimated using the Cohen Kappa (κ) index [10–13]. The significance level α was set to 0.05. Each variable was coded binary. The null hypothesis for agreement is a κ of 0.

3. Results

One hundred patients were included between January 2013 and April 2014. The progress through the fases of this inter-rater reliability study is presented in Fig. 2. Demographic data of the patients were a median age of 55 (interquartile range (27,25) 65.75–44.25), a mean BMI of 26.8 (standard deviation 5.6), 66% female gender and 100% Caucasian race.

Inter-rater reliability (Kappa (κ) measure of agreement) and significance (p) between raters (raters A, B and C are pain physicians (two physicians for each patient), rater D an orthopedic surgeon) are presented in Tables 2a–c. Strengths of agreement are presented in order of agreement for values $\kappa > 0.20$ in Table 3. None of the parameters from the physical investigation had κ values of more than 0.21 (fair) in all pairs of raters. Between two raters (C and D), there was an almost perfect agreement on three parameters, more specifically “Abnormal sensory and motor examination, hyperactive or diminished reflexes”, “Sitting

Table 2a

Inter-rater reliability (Kappa measure of agreement) and significance (p) between raters (raters A, B and C are pain physicians, rater D and orthopedic surgeon) of the physical examination suggestive of SI joint pain. 1: Drop-test positive; 2: Sitting exam shows no reflex, motor or sensory signs in the legs; 3: Straight leg raising (Laségue) negative between 30 and 70 degrees of passive flexion; 4: Distraction (Gapping) test positive; 5: Posterior shear (thigh trust) test positive; 6: Pelvic torsion (Gaenslen's) test positive; 7: Patrick-Faber test positive; 8: Compression test positive; 9: Sacral thrust test positive; 10: Cranial shear test positive; 11: Bilateral internal rotation of the hip/Unilateral rotation of the hip painful at SI joint(s); 12: Yeoman's test positive

Nr.	A-B	A-C	A-D	B-C	B-D	C-D
1	0.23 (0.01)	−0.02 (0.45)	0.43 (0.03)	0.06 (0.31)	0.00 (1.00)	0.14 (0.10)
2	−0.40 (0.00)	−0.28 (0.00)	0.03 (0.00)	0.10 (0.01)	−0.22 (0.00)	0.86 (0.00)
3	−0.19 (0.04)	−0.02 (0.55)	0.21 (0.00)	0.00 (−)	−0.28 (0.00)	0.81 (0.00)
4	−0.02 (0.69)	−0.01 (0.50)	−0.04 (0.02)	0.00 (1.00)	0.03 (0.31)	−0.14 (0.07)
5	−0.02 (0.79)	−0.06 (0.08)	−0.03 (0.15)	−0.04 (0.49)	0.02 (0.55)	−0.02 (0.85)
6	−0.09 (0.29)	−0.03 (0.28)	−0.11 (0.00)	−0.02 (0.74)	0.07 (0.03)	−0.28 (0.00)
7	−0.07 (0.51)	−0.06 (0.10)	−0.09 (0.00)	−0.07 (0.20)	0.04 (0.21)	−0.21 (0.02)
8	0.02 (0.69)	−0.02 (0.36)	−0.03 (0.09)	0.06 (0.26)	0.06 (0.03)	0.00 (1.00)
9	−0.16 (0.12)	−0.04 (0.21)	0.01 (0.59)	0.02 (0.75)	−0.08 (0.02)	0.12 (0.17)
10	−0.02 (0.81)	0.02 (0.38)	−0.01 (0.59)	−0.10 (0.06)	−0.04 (0.18)	−0.10 (0.20)
11	0.07 (0.42)	0.01 (0.79)	0.04 (0.01)	0.04 (0.43)	−0.01 (0.80)	0.10 (0.15)
12	0.02 (0.81)	0.02 (0.63)	0.02 (0.50)	−0.09 (0.08)	0.02 (0.56)	−0.10 (0.21)

Table 2b

Inter-rater reliability (Kappa measure of agreement) and significance (p) between raters (raters A, B and C are pain physicians, rater D an orthopedic surgeon) of the physical examination suggestive of disc pain. 1: Gait deviation; 2: Abnormal sensory and motor examination, hyperactive or diminished reflexes; 3: Digital Interspinous Pressure (DIP) test positive

Nr.	A-B	A-C	A-D	B-C	B-D	C-D
1	0.09 (0.14)	−0.03 (0.09)	0.02 (0.09)	0.02 (0.31)	0.00 (1.00)	0.10 (0.01)
2	−0.09 (0.31)	−0.26 (0.00)	0.04 (0.00)	0.20 (0.00)	−0.12 (0.00)	0.91 (0.00)
3	0.30 (0.00)	−0.03 (0.48)	0.10 (0.00)	0.22 (0.00)	0.01 (0.78)	0.42 (0.00)

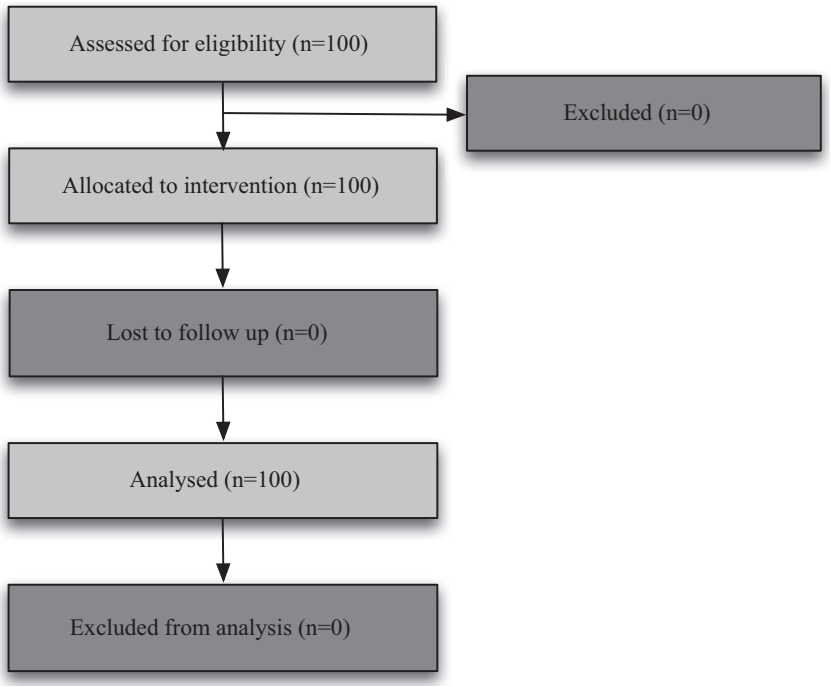

Fig. 2. Flow diagram of the progress through the phases of the inter-rater reliability study.

Table 2c
Inter-rater reliability (Kappa measure of agreement) and significance (p) between raters (raters A, B and C are pain physicians, rater D an orthopedic surgeon) of the physical examination suggestive of facet joint pain. 1: Straight leg raising (Laségue) positive between 30 and 70 degrees of passive flexion; 2: Pain in extension; 3: Pain eased in flexion; 4: Pain when rising from forward flexion; 5: Schober test $<$ 3–5 cm; 6: Pain in extension, lateral flexion or rotation manoeuvers to the ipsilateral side; 7: Replication or aggrevation of pain by unilateral pressure over the ipsilateral side; 8: Local unilateral passive movements show reduced range of motion or increased stiffness on the side of the involved facet joints; 9: Tight or facilitated muscles (psoas, hip adductors, gluteus medius muscles); 10: Weak muscles (gluteus maximus, gluteus medius)

Nr.	A-B	A-C	A-D	B-C	B-D	C-D
1	0.07 (0.29)	−0.01 (0.50)	0.02 (0.18)	0.00 (−)	0.00 (1.00)	0.04 (0.51)
2	−0.05 (0.66)	0.04 (0.34)	−0.10 (0.00)	0.00 (1.00)	0.08 (0.00)	−0.30 (0.00)
3	−0.07 (0.29)	−0.01 (0.69)	−0.14 (0.00)	0.09 (0.08)	0.17 (0.00)	−0.39 (0.00)
4	−0.09 (0.38)	−0.07 (0.02)	−0.09 (0.00)	0.00 (1.00)	0.00 (0.86)	−0.04 (0.71)
5	0.19 (0.02)	−0.04 (0.08)	−0.08 (0.00)	0.07 (0.15)	0.14 (0.00)	−0.18 (0.04)
6	0.07 (0.46)	0.02 (0.60)	−0.06 (0.01)	−0.04 (0.31)	0.07 (0.00)	−0.21 (0.01)
7	0.00 (1.00)	−0.12 (0.00)	−0.13 (0.00)	0.06 (0.20)	0.09 (0.00)	0.00 (1.00)
8	0.16 (0.09)	−0.03 (0.33)	−0.19 (0.00)	0.07 (0.18)	0.21 (0.00)	−0.37 (0.00)
9	0.12 (0.11)	0.02 (0.59)	0.05 (0.00)	0.07 (0.07)	0.01 (0.70)	0.10 (0.11)
10	−0.16 (0.10)	−0.01 (0.79)	0.06 (0.00)	−0.06 (0.21)	−0.15 (0.00)	0.21 (0.00)

Table 3
Strength of agreement beyond chance, indicated with κ values above 0.20 ($<$ 0: poor; 0–0.20: slight; 0.21–0.40: fair; 0.41–0.60: moderate; 0.61–0.80: substantial; 0.81–1.00: almost perfect). The κ values used are from Landis and Koch [12] and are in order in agreement

κ	Nr	Raters
0.91	Abnormal sensory and motor examination, hyperactive or diminished reflexes	C-D
0.86	Sitting exam shows no reflex, motor or sensory signs in the legs	C-D
0.81	Straight leg raising (Laségue) negative between 30 and 70 degress of passive flexion	C-D
0.43	Drop test positive	A-D
0.42	Digital interspinous pressure test positive	C-D
0.30	Digital interspinous pressure test positive	A-B
0.23	Drop test positive	A-B
0.22	Digital interspinous pressure test positive	B-C
0.21	Weak muscles (gluteus maximus, gluteus medius)	C-D
0.21	Local unilateral passive movements show reduced range of motion or increased stiffness on the side of the involved facet joints	B-D
0.21	Straight leg raising (Laségue) negative between 30 and 70 degrees of passive flexion	A-D

exam shows no reflex, motor or sensory signs in the legs" and "Straight leg raising (Laségue) negative between 30 and 70 degrees of flexion". The "Drop test positive" parameters had moderate strength of agreement between raters A and D and fair strength between raters A and B. The "Digital interspinous pressure test positive" had moderate strength of agreement between raters C and D and fair strength of agreement between raters A and B as well as raters B and C. Three other parameters (Table 3) had a fair strength of agreement between two raters, all other parameters had a slight or poor strength of agreement.

Inter-rater reliability (including confidence intervals and significance) of pooled items for SI joint-, disc- and facet joint pain are represented in Tables 4a-c. Kappa values for the pooled parameters of the physical examination suggestive of SI joint pain stayed below 0.2 between all raters. The same applies for the pooled parameters of the physical examination suggestive of facet joint- or disc pain.

During the study we recorded no (serious) adverse events.

4. Conclusion and discussion

We conducted this study to determine the reliability of diagnostic tests that point towards SI joint-, disc- or facet joint pain, using diagnostic tests mentioned in the literature on these subjects. The null hypothesis for agreement is a κ of 0. None of the diagnostic tests used in this study had κ values of more than 0.21 (fair) in all pairs of raters. Also, the κ values in all pairs of raters of the pooled items of the physical examination parameters suggestive for SI joint-, disc- or facet joint pain stayed below 0.2. The poor reliability of the diagnostic parameters seriously limits their predictive validity, and as such their use in patients with low back pain for more than 3 months.

Kappa is an adequate measure for inter-rater agreement. Kappa has the advantage that it is corrected for

Table 4a
Inter-rater reliability (Kappa measure of agreement) (raters A, B and C are pain physicians, rater D an orthopedic surgeon), significance (p) and 95% confidence intervals (CI) of the pooled items of the physical examination parameters suggestive for SI joint pain

	Rater A κ;p; (95% CI κ)	Rater B κ;p; (95% CI κ)	Rater D κ;p; (95% CI κ)
Rater A			0.124; 0.006 (0.034; 0.214)
Rater B	0.169; $<$ 0.001 (0.085; 0.252)		0.136; 0.001 (0.052; 0.219)
Rater C	0.130; 0.004 (0.035; 0.225)	0.166; $<$ 0.001 (0.082; 0.251)	0.036; 0.44 (0; 0.129)

Table 4b
Inter-rater reliability (Kappa measure of agreement) (raters A, B and C are pain physicians, rater D an orthopedic surgeon), significance (p) and 95% confidence intervals (CI) of the pooled items of the physical examination parameters suggestive for disc pain (N/A: concordance is smaller than mean-chance)

	Rater A κ;p; (95% CI κ)	Rater B κ;p; (95% CI κ)	Rater D κ;p; (95% CI κ)
Rater A			0.194; 0.000 (0.075; 0.313)
Rater B	N/A		0.191; 0.001 (0.009; 0.373)
Rater C	0.205; 0.003 (0.070; 0.341)	0.093; 0.145 (0; 0.232)	0.129; 0.001 (0.051; 0.207)

Table 4c
Inter-rater reliability (Kappa measure of agreement) (raters A, B and C are pain physicians, rater D an orthopedic surgeon), significance (p) and 95% confidence intervals (CI) of the pooled items of the physical examination parameters suggestive for facet joint pain

	Rater A κ;p; (95% CI κ)	Rater B κ;p; (95% CI κ)	Rater D κ;p; (95% CI κ)
Rater A			0.313; 0.000 (0.255; 0.372)
Rater B	0.258; 0.000 (0.173; 0.343)		0.307; 0.000 (0.241; 0.373)
Rater C	0.357; 0.000 (0.275; 0.440)	0.232; 0.000 (0.111; 0.354)	0.276; 0.000 (0.205; 0.346)

agreement with statistical chance. The main disadvantage is that it is not free of dependence on disease prevalence or the number of rating categories. As a consequence it can be difficult to interpret the meaning of any absolute value, but is still useful if disease prevalence and number of categories are presented [12].

In correlating the clinical examination characteristics in 81 individuals (a total of 104 injection procedures were performed), both centralization of pain and pain when rising from sitting were significantly associated with a positive discogram [2], while not having pain when rising from sitting was strongly correlated with a positive facet joint injection. The presence of midline lumbar pain tends to exclude the SI joint as a potential pain generator. When there were three or more positive SI joint pain provocation tests, the presence of a SI joint source of pain is 28 times more likely. The physical examinations were performed by visiting physical therapists and the injections were performed if requested by the referring physician or deemed adequate by a radiologist, while in our study all parts of the trial were performed by the same physicians and on the basis of a general working diagnosis.

In a systematic review of tests to identify the source of low back pain, no available clinical test was found which could be used to increase or decrease the likelihoofd of the disc as the source of low back pain [3]. Also, the currently available tests have limited or no diagnostic validity regarding investigating the facet joint as the source of low back pain; our study is in accordance with this review in that we also found no useful diagnostic tests.

A combination of SI joint provocation tests appears to be useful to increase the likelihood of the SI joint as the source of pain. However, in a small study performed by physical therapists examining the intertester reliability of tests for SI joint dysfunction, the reliability was poor for all tests, except the iliac gapping and compression tests [14]. In our study, we found that no single parameter of the physical examination nor the pooling of these tests was useful to increase the likelihood of the SI joint as the source of pain; the same applies to the parameters of the physical examination suggestive for disc – or facet joint pain.

Only a small amount of investigation has been performed into the diagnostic accuracy of clinical tests. In our study we investigated the diagnostic accuracy of these tests in 100 patients referred to a pain center because of chronic low back pain and found a poor reliability of all diagnostic parameters.

Acknowledgements

Fleur A. Schuurmans, RN (Registered nurse), David C. van den Tol, MD, FIPP (Consultant Anesthesiologist and Pain Specialist), Jacqueline van Vliet, MD (Consultant Anesthesiologist and Pain Special-

ist) and Geert Meermans, MD (Consultant Orthopedic Surgeon) of Bravis Hospital, Bergen op Zoom, The Netherlands.

Conflict of interest

All authors declare that no support from any organisation for the submitted work has been received, no financial relationships with any organisations have been established that might have an interest in the submitted work and no other relationships or activities were established that could appear to have influenced the submitted work.

References

[1] Robinson HS, Brox JI, Robinson R, Bjelland E, Solem S, Telje T. The reliability of selected motion- and pain provocation tests for the sacroiliac joint. Manual Therapy 2007; 12: 72-9.

[2] Young S, Aprill C, Laslett M. Correlation of clinical examination characteristics with three sources of chronic low back pain. The Spine Journal 2003; 3: 460-5.

[3] Hancock MJ, Maher CG, Latimer J, Spindler MF, McAuley JH, Laslett M, et al. Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. Eur Spine J 2007; 16: 1539-50.

[4] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 2011; 64: 96-106.

[5] Slipman CW, Sterenfeld EB, Chou LH, Herzog R, Vresilovic E. The predictive value of provocative sacroiliac joint stress maneuvres in the diagnosis of sacroiliac joint syndrome. Arch Phys Med Rehabil 1998; 79: 288-92.

[6] Robinson HS, Brox JI, Robinson R, Bjelland E, Solem S, Telje T. The reliability of selected motion- and pain provocation tests for the sacroiliac joint. Manual therapy 2007; 12: 72-9.

[7] Laslett M, Williams M. The reliability of selected pain provocation tests for sacroiliac joint pathology. SPINE 1994; 19: 1243-9.

[8] 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.

[9] Laslett M, Aprill CN, McDonald B, Young SB. Diagnosis of sacroiliac joint pain: validity of individual provocation tests and composites of tests. Manual Therapy 2005; 10: 207-18.

[10] Kundel HL, Polansky M. Measurement of Observer Agreement. Radiology 2003; 228: 303-8.

[11] Siegel TS, Castellan Jr NJ. Nonparametric statistics for the behavioral sciences, second edition, New York: McGraw-Hill, 1988.

[12] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37-46.

[13] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1997; 33: 159-74.

[14] Potter NA, Rothstein JM. Intertester reliability for selected clinical tests of the sacroiliac joint. Phys Ther 1985; 65: 1671-5.