

Original article

Clinical tests of the sacroiliac joint

A systematic methodological review. Part 2: Validity

P. van der Wurff*, W. Meyne†, R.H.M. Hagmeijer*

*Manual Therapist, Department of Physiotherapy, Military Rehabilitation Centre 'Aardenburg', Doorn, The Netherlands, †Physiotherapist, Private Practice, Bussum, The Netherlands

SUMMARY. In the literature many tests are described which are designed to provoke pain or detect joint mobility in the sacroiliac joint (SIJ). However, in part 1 of this review, the authors stated that there is little evidence of reliability of these tests. In this article, the authors describe the methodological review of 11 studies, which have dealt with the validity of SIJ tests. The methodological quality of the studies was tested by using a list of criteria that consisted of three categories: 1) study population, 2) test procedure and 3) test results. A weighting for each criterion was developed. The methodological score for the studies was, in general, disappointing and looked promising for only two out of 11 studies (58 and 64 points). Four authors drew conclusions of positive validity from the tests they studied but other authors did not confirm these results. The conclusion of this methodological review is that there is no evidence to support the inclusion of mobility and pain provocation tests for the SIJ in clinical practice. Three major problems have been identified in validating SIJ dysfunction tests. Firstly, poor reliability of SIJ dysfunction tests exists, which may be improved by multiple test scores as postulated in part 1 of this review. Secondly, the methodological quality of validity studies needs to be developed to a much higher level with special consideration paid to sensitivity, specificity, confidence intervals and likelihood ratio values. And finally, there is a need for the proper use of a gold standard in assessing the validity of SIJ tests. © 2000 Harcourt Publishers Ltd.

INTRODUCTION

In part 1 of this systematic review of the methodological quality of studies concerning the value of clinical tests for the sacroiliac joint (SIJ), the authors described and critiqued reliability studies for SIJ tests (Van der Wurff et al. 1999).

During the literature search for part 1, a number of articles had to be rejected because they did not conform to the inclusion criteria of examining reliability of pain provocation tests or mobility tests for the SIJ. Other articles were rejected in the methodological review of the reliability of the SIJ tests because authors had accepted tests for the SIJ on authority-based evidence as a reliable and valid method in screening selected populations or to verify various treatment procedures (Miereau et al. 1984;

Gajdosik et al. 1985; Burdett et al. 1986; Alviso et al. 1988; Cibulka et al. 1988; Gemmel & Jacobson 1990; Dreyfuss et al. 1994).

Some articles were also rejected for use in part 1 of this study if they dealt with validity rather than reliability (Russell et al. 1981; Blower & Griffin 1984; Bemis & Monte 1987; Rantanen & Airaksinen 1989; Östgaard et al. 1994; Maigne et al. 1996; Egan et al. 1996; Dreyfuss et al. 1996). However both valid and reliable test results are necessary, as reliability alone is insufficient to underpin the quality of a diagnostic test.

The validity of a test procedure is defined as the ability of a test to determine the presence or absence of the pathological condition correctly (Jaeschke et al. 1994a, 1994b). In health care there is a preference for diagnostic tests which are fast, simple, inexpensive and as minimally invasive as possible for the patient. To assess the value of a diagnostic test it is necessary to compare the results with those obtained from a 'reference' test (the so-called 'gold standard'). The 'gold standard' purports to be the test with more or less 100% validity. Very often the 'gold

Peter van der Wurff, Ruud H.H. Hagmeijer, Department of Physiotherapy, Military Rehabilitation Centre 'Aardenburg', Postbus 185, 3941 AD Doorn, Wilco Meyne, Physiotherapist in Private Practice, Bussum, The Netherlands.

Correspondence to PW.

standard test' consists of a procedure, which is expensive, time consuming, laborious, and in some cases, leads to some risks for the patient. Initially, the 'gold standard' related to histological evidence. Later, the term 'gold standard' was widely used to describe the 'best', or most appropriate, way to confirm a diagnosis. In some cases, the validity of a diagnostic test would be assessed by comparing the results of the test with the results of an external criterion (criterion validity as a reference test). Validity is directly related to the notion of sensitivity and specificity. Sensitivity represents the proportion in a population with the pathologic condition, who have had a positive result from the diagnostic test. Specificity is the proportion of a study population without the pathologic condition in whom the test result is negative. Sensitivity and specificity are directly related to each other and a test with an outstanding sensitivity but low specificity, and vice versa, is of little value. The acceptable values for sensitivity and specificity depend on the context of the subject that is studied. They are generally positioned between 50% (unacceptable test) and 100% (perfect test), the arbitrary cut-off point is 80% on authority-based evidence for studies in manual therapy. Other points are the prediction value, confidence interval and likelihood ratio of a diagnostic test. The prediction value of a positive test indicates that those members of the study population who have a positive test outcome will suffer from the condition under investigation. The diagnostic power of the negative test outcome relates to those of the study population with a negative test outcome, who do not suffer from the condition under investigation. The 95% confidence interval expresses the limits within which the actual association of 95% acceptance is positioned (Jaeschke et al. 1994b).

The likelihood ratio is the index-measurement, which combines sensitivity and specificity values. The likelihood ratio indicates how much a given diagnostic test result will raise or lower the pre-test probability of the target disorder (Jaeschke et al. 1994a, 1994b).

For some years, it has been more or less accepted that an anaesthetic block procedure for SIJ dysfunction could be used as the gold standard for diagnostic tests (Dreyfuss et al. 1996; Maigne et al. 1996; Fortin & Falco 1997; Slipman et al. 1998; Broadhurst & Bond 1998). The validity of the mobility tests is another point of discussion. Mobility tests differ from pain provocation tests because constructing the 'gold or reference' standard is even more difficult. In fact, the only options are in vivo studies, (Sturesson 1998) or the use of a three-dimensional recording system (Hungerford & Gilleard 1998). Sturesson (1998) studied the Gillet-test with the invasive radiostereometric method using 22 patients with presumed SIJ dysfunction and a positive Gillet test. The results

showed no significant difference between the left and right SIJ and the author concluded that there was no argument for using this test. Hungerford and Gilleard (1998) used a non-invasive procedure with skin markers and a five camera recording system to assess SIJ mobility and the effects of the Gillet test. Their conclusion was positive and supported the use of the Gillet test. These studies, however, were not fully reported since they contained insufficient data, and therefore they are not included in the present systematic review.

In Part 1 of this methodological review, the authors stated that the reliability results of pain provocation tests and mobility tests of the SIJ were not convincing and therefore do not justify further validity studies at this time. Nevertheless, in this article, the authors will present a systematic methodological review of previous validity studies of SIJ tests for the sake of completeness.

The aim of this systematic methodological review of the validity of SIJ pain provocation and mobility tests was to determine the validity of tests that are relevant in daily clinical practice. In addition, the authors aim to make suggestions for further research in this area.

METHODS

Study selection

For this systematic methodological review, the authors included studies that met the following conditions:

- The results were published as a full report before February 1999 (Abstracts, letters and unpublished studies were not selected)
- All relevant clinical tests of the sacroiliac joint were included
- The studies were to be studies of validity
- Studies were to be written in English, German, French or Dutch.

A Medline, Embase and CINAHL literature search was carried out from the period between January 1980 and February 1999. The keywords used were: sacroiliac joint, physical examination, palpation, evaluation studies, validity and assessment. In addition, the references found in selected publications were also examined.

Methods assessment

Studies were identified by PW by computer search. From the potential relevant publications, a total of eight were selected as suitable for inclusion in this study. A further three studies were identified from references cited in selected articles. The selected publications were blinded for author(s), source of publication, results and conclusions in order to

minimize potential reviewer bias. RHMH and WM independently scored each publication according to a standardized set of 20 methodological criteria (Table 1).

The development of the criteria list is fully described in the methodological review in part 1 of this article (Van der Wurff et al. 1999). The authors adopted the guidelines for meta-analysis-evaluating diagnostic tests by Irwig et al. (1994) and Reid et al. (1995) and the methodological guidelines for systematic reviews by Van Tulder et al. (1997). Items that seemed to be irrelevant for validity studies were dropped and more appropriate items were added. In comparison with part 1 of this study, the authors removed the test/re-test item as well as the inferential statistics (Cohen's Kappa). These items were replaced by the 'gold standard' or 'reference standard' (item G) and data concerning sensitivity and specificity respectively (item J). The authors chose data concerning sensitivity and specificity as a minimum criterion. We did not include confidence interval and likelihood ratios for this item since these items were relatively unfamiliar until a few years ago and therefore were not expected to be used in the studies included in this review.

Statistical analysis

Statistical analysis of the data was performed using SPSS version 8.0. The Cohen's Kappa (κ) was used to

obtain the inter-examiner reliability for the reviewers of the selected articles. The authors of this review calculated values for sensitivity, specificity and likelihood ratio if sufficient data was available and if the original authors had not included this aspect of data analysis.

RESULTS

Eleven articles met the inclusion criteria (Russell et al. 1981; Blower & Griffin 1984; Bemis & Monte 1987; Rantanen & Airaksinen 1989; Östgaard et al. 1994; Maigne et al. 1996; Egan et al. 1996; Dreyfuss et al. 1996; Fortin & Falco 1997; Slipman et al. 1998; Broadhurst & Bond 1998). One article (Dreyfuss et al. 1996) was selected for the reliability review (Part 1) as well as the validity review because both aspects were described separately.

The 11 studies are presented in Table 2 in a hierarchical order according to their methodological score. Initially in judging the studies, the two reviewers failed to agree in 14 out of 220 instances using the criterion (percentage disagreement 6%, CI-interval 4.1–10.1).

In the majority of cases, this appeared to be due to an error in reading. The disagreement between the two reviewers occurred mostly with the items: description of inclusion and exclusion criteria, gold

Table 1. Criteria list for methodological assessment of validity trials for SIJ-dysfunction

Criteria	Weighting
Study population	
A 1. Description of study population i.e. volunteers or patients, age, gender etc.	8
2. Description inclusion and exclusion criteria	7
B Drop-outs described, information from which group and with reason for withdrawal	5
C Number of subjects	
< 25 subjects	0
> 25 subjects	3
> 50 subjects	6
> 75 subjects	10
Test procedure	
D Standardization of test procedure	
1. Position of subject	3
2. Position of examiner	2
3. Description palpation technique (position hands of examiner)	3
4. Description neutralising simple exercises for low back and pelvis before or during the test procedure	2
5. Information given to the subject about the test procedure	2
6. Standardisation according to the original description of the test in the literature (referenced)	4
E Selection of examiner	
1. Description of the choice for experienced examiners	3
2. Description of less-experienced examiner	2
3. Description of a consensus procedure	9
Test results	
F Standardized measurement of test outcome	
G Description of gold standard or reference standard	5
H Procedure of blinding	2
1. Attempt of blinding the examiner	
2. Subject not informed of outcome	2
3. Results sealed: the examiners could not see each others' findings	1
I Descriptive statistics: frequencies	5
J Inferential statistics: data about sensitivity and specificity	10
	15

The answer options are yes or no.

or reference standard and the inferential statistics. After a consensus meeting between the two reviewers, the disagreements were completely resolved and it was not necessary for a third reviewer to make a final decision.

The inter-examiner reliability between the two reviewers was found to be $\kappa=0.82$, which is 'almost perfect' agreement according to the classification of Landis and Koch (1977). A comparison of the results by the reviewers with the results of part 1 of this study and with similar studies, showed relatively equal results (Koes 1994). In this review, only two studies (Östgaard et al. 1994; Dreyfuss et al. 1996) gained methodological scores of more than 51 points and nine scored less than 51 points. These low methodological scores generally indicate the poor quality of the studies included.

Table 2 shows that the most prevalent methodological problems concerned:

- the drop-outs described (B);
- the description of neutralising movement during the test (D4);
- information given to the subjects (D5);
- the selection of the examiner (E1-3);
- blinding of the examiner and patient (H1-2);
- results sealed (H-3)
- inferential statistics (J).

A small majority of the studies (six studies) reported negative results for validity of SIJ tests for joint mobility and pain provocation tests (Russell et al. 1981; Rantanen & Airaksinen 1989; Maigne et al. 1996; Egan et al. 1996; Dreyfuss et al. 1996; Slipman et al. 1998). In five studies positive results for validity were described (Blower & Griffin 1984; Bemis & Monte 1987; Östgaard et al. 1994; Fortin & Falco 1997; Broadhurst & Bond 1998). The methodological score for the Fortin and Falco (1997) study was the lowest and the Östgaard study the highest. Both studies were carried out during the past 4 years.

Table 2 shows the overall review of validity studies for the pain provocation tests and mobility tests of the SIJ in order of methodological score. The majority of the studies reported satisfactory homo-

geneity (item A1); only one (Bemis & Monte 1987) failed to do so. The drop-out rates were described in two studies (Östgaard et al. 1994; Dreyfuss et al. 1996) and partly by Maigne et al. (1996). In the other studies, this item was not mentioned. Compared with the data from the reliability review the description of the investigated tests (item D) was less extensive and showed several shortcomings. The selection of the observers was only mentioned by Egan et al. (1996). The 'gold or reference standard' was described in eight cases: three used a reference standard (Russell et al. 1981; Blower & Griffin 1984; Rantanen & Airaksinen 1989) and five used anaesthetic arthrogram blocks as a 'gold standard' (Maigne et al. 1996; Dreyfuss et al. 1996; Fortin & Falco 1997; Slipman et al. 1998; Broadhurst & Bond 1998).

An attempt at blinding the patient and examiner was only performed by Östgaard et al. (1994). With respect to the presentation of data concerning the sensitivity and specificity (item J), only three studies considered this item (Östgaard et al. 1994; Dreyfuss et al. 1996; Broadhurst & Bond 1998).

In the three remaining studies, it was possible to calculate sensitivity and specificity rates because sufficient data was presented in the published papers (Russell et al. 1981; Blower & Griffin 1984; Rantanen & Airaksinen 1989).

Pain provocation tests

Table 3 shows the validity of the pain provocation tests for the SIJ by author. As stated before, only three studies presented data on sensitivity and specificity (Östgaard et al. 1994; Dreyfuss et al. 1996; Broadhurst & Bond 1998). In three studies, the data was presented in the original paper in such a way that the authors were able to calculate values for sensitivity and specificity but not for likelihood ratios (Russell et al. 1981; Blower & Griffin 1984; Rantanen & Airaksinen 1989). Maigne et al. (1996), Fortin and Falco (1997) and Slipman et al. (1998) did not describe appropriate frequencies and for that reason it was impossible to calculate specificity and/or sensitivity values. Maigne et al. (1996) used the χ^2

Table 2. Overall results of the validity of pain provocation tests and mobility tests for the SIJ in order of methodological score

First author	Year	A1	A2	B	C	D1	D2	D3	D4	D5	D6	E1	E2	E3	F	G	H-1	H-2	H-3	I	J	Method score
		8	7	5	10	3	2	3	2	2	4	3	2	9	5	10	2	1	5	2	15	
Östgaard	1994	8	7	5	6	3	2	3	0	0	0	0	0	0	5	0	2	1	5	2	15	64
Dreyfuss	1996	8	0	5	10	3	0	0	0	0	0	0	0	0	5	10	0	0	0	2	15	58
Broadhurst	1998	8	7	0	3	0	0	0	0	0	0	0	0	0	5	10	0	0	0	2	15	50
Rantanen	1989	8	7	0	3	3	2	3	0	0	4	0	0	0	5	10	0	0	0	0	0	45
Blower	1984	8	7	0	10	3	0	0	0	0	0	0	0	0	5	10	0	0	0	2	0	45
Egan	1996	8	7	0	10	0	0	0	0	0	4	3	2	0	5	0	0	0	0	2	0	41
Russell	1981	8	7	0	10	0	0	0	0	2	0	0	0	0	0	10	0	0	0	2	0	39
Maigne	1996	8	7	3	6	3	0	0	2	0	0	0	0	0	0	10	0	0	0	2	0	39
Slipman	1998	8	7	0	3	0	0	0	0	0	4	0	0	0	0	10	0	0	0	0	0	32
Bemis	1987	0	0	0	6	3	2	3	0	2	0	0	0	0	5	0	0	0	0	2	0	23
Fortin	1997	8	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	18

Table 3. Overview of the validity by author for individual SIJ pain provocation tests

Test	First author	Sensitivity	Specificity	Method score	Authors' conclusion
Gapping or distraction test	Rantanen	15%*	na	45	Not valid
	Blower	21%*	100%*	45	Valid
	Russell	11%*	90%*	39	Not valid
	Maigne	na	na	26	Not valid
Compression test	Rantanen	19%*	na	45	Not valid
	Blower	0%*	100%*	45	Not valid
	Russell	7%*	90%*	39	Not valid
	Maigne	na	na	39	Not valid
Gaenslen test	Dreyfuss	71%	26%	58	Not valid
	Russell	21%	72%*	39	Not valid
	Maigne	na	na	39	Not valid
Sacral thrust	Dreyfuss	53%	29%	58	Not valid
	Blower	27%	100%	45	Valid
	Russell	3%*	90%*	39	Not valid
	Maigne	na	na	26	Not valid
Thigh thrust	Östgaard	80%	81%	64	Valid
	Dreyfuss	36%	50%	58	Not valid
	Broadhurst	80%	100%	50	Valid
Patrick's sign	Dreyfuss	69%	16%	58	Not valid
	Rantanen	57%*	na	45	Not valid
	Maigne	na	na	39	Not valid
	Broadhurst	77%	100%	50	Valid
Resisted abduction	Broadhurst	77%	100%	50	Valid
Extension of the hip	Russell	21%*	72%*	39	Not valid
Mennell's sign	Rantanen	50%*	na	45	Not valid
Fortin finger test	Fortin	na	na	18	Valid
MTS [†]	Slipman	na	na	32	Not valid

na = not available; [†]multiple-test score; *calculated by the authors of this article.

test and Yates test and concluded that there was no association between pain provocation tests and the response to anaesthetic blocks. Because of the minimal amount of raw data they described, it was not possible for the present authors to calculate the sensitivity and specificity rates.

For the gapping or distraction tests, only Blower and Griffin (1984) reported a positive conclusion based on inappropriate statistics (χ^2 test). Their method score was low, less than 50 points and consisted of high specificity but low sensitivity. Russell et al. (1981) reported similar results but their conclusion was negative. Rantanen and Airaksinen (1989) and Maigne et al. (1996) also reported a negative result for validity based on their frequencies. All four studies using the compression test led to a negative conclusion with three of them producing low sensitivity and two of them acceptable specificity. The Gaenslen test was studied in three of the studies and all the authors concluded that the results of validity tests were negative (Russell et al. 1981; Dreyfuss et al. 1996; Maigne et al. 1996).

The Sacral thrust test was studied in four of the papers. Only Blower and Griffin (1984) concluded positive results for validity, yet these appeared to be based on inappropriate statistics. The Thigh thrust test was studied on three occasions: Dreyfuss et al. (1996) concluded that their results of validity testing

were negative. Östgaard et al. (1994) and Broadhurst and Bond (1998) reported acceptable levels of sensitivity and specificity and stated positive results for validity tests. Östgaard et al. (1994) calculated a positive prediction value of 71% and a negative prediction value of 88%. The present authors calculated the likelihood ratios for a positive test as 4.07 and for a negative test result as 0.23. These values suggested a small change in pre- and post-test probability.

Dreyfuss et al. (1996) reported Patrick's sign as having a (relatively) high score for sensitivity (69%), which was unacceptable according to our standards, but they also reported low specificity and therefore, negative results for validity. Rantanen and Airaksinen (1989) and Maigne et al. (1996) also concluded negative results for validity. Broadhurst and Bond (1998) reported positive results for validity in their study of Patrick's sign. Studies of Mennell's sign (Rantanen & Airaksinen 1989) and extension of the hip (Russell et al. 1981) scored negatively. Fortin and Falco (1997) described positive test results for the Fortin finger test as did Broadhurst and Bond (1998) for the Resisted abduction test. However, both tests scored low methodologically. Broadhurst & Bond (1998) used an inappropriate study design (with, in particular, the absence of a control group) and did not draw any conclusions about specificity and

sensitivity values for the pain provocation tests they studied. Slipman et al. (1998) provided the only (weighted) multiple test score consisting of three positive tests: Patrick's sign and palpation pain over the sacral sulcus and in addition one of the following tests: Shear test, Standing extension test, Gaenslen's test or the Yeoman test. Their conclusions, however, were negative.

Mobility tests

Table 4 shows the validity of the mobility tests for the SIJ. Egan et al. (1996) compared the Overtake phenomenon, which they described as the Forward Flexion test, with a reference standard constructed by factors such as low back pain, pelvic asymmetry, age, height, weight and standing symmetry. They concluded that factors other than SIJ dysfunction are likely to be the cause of a positive Overtake phenomenon.

Dreyfuss et al. (1996) used the anaesthetic block as the gold standard and compared it with the Gillet test for detecting mobility of the SIJ. They reported low sensitivity (43%) and a higher specificity (68%). Both values, however, were unacceptable.

Bemis and Monte (1987) constructed a reference standard for a group of subjects with SIJ dysfunction, which had a combination of a positive Standing flexion test, a negative Sitting flexion test and an asymmetric Posterior Superior Iliac Spine (PSIS) level. The control group had a negative Standing flexion test, a negative Sitting flexion test and a symmetric PSIS level.

Bemis & Monte (1987) used the χ^2 test and concluded that there was a significant difference between the SIJ dysfunction group and the control group. The authors of this present article calculated the sensitivity and specificity values for the study but both values were found to be unacceptable.

DISCUSSION

In this present review, the authors evaluated the methodological quality of 11 studies in which the original authors had assessed validity of mobility tests and pain provocation tests for the SIJ.

One of the major problems in attempting to validate pain provocation and mobility tests of the SIJ dysfunction is the absence of morphological data

as demonstrated, for example, by radiographs or laboratory results. Accepting an anaesthetic block as a gold standard is questionable, as stated by Laslett (1998) and Tanner (1997). Firstly, the anaesthetic block arthrogram seems to investigate intra-articular sources of pain and not the whole SIJ complex including the ligaments. Secondly, there is no absolute certainty about the possibility of anaesthetic block affecting all parts of the joint capsule.

Another point of discussion directly related to this gold standard is the use of the visual analogue scale (VAS), in particular the cut-off point for discriminating whether an anaesthetic block procedure is successful or not. For example, Dreyfuss et al. (1996) used a cut-off point of 90% relief of pain: Slipman et al. (1998) 80%; Maigne et al. (1996) 75% and Broadhurst and Bond (1998) 70%. The final conclusions as postulated by Broadhurst and Bond (1998), can be challenged significantly. In their study, the positive outcome decreased by 50% if the cut-off point was revalued from 70 to 90%.

When using an anaesthetic injection one can expect the VAS score to be 100% if the procedure is successful and in the authors' opinion it is advisable to choose the cut-off point for the VAS as close to 100% as possible and at least 90% as advised by Dreyfuss et al. (1996).

Some authors (Russell et al. 1981; Blower & Griffin 1984; Rantanen & Airaksinen 1989) used the pathological condition of their study population, i.e. ankylosing spondylitis (AS), as a reference standard. This choice is debatable because it is not known how long the patients in these studies had suffered from AS. It is possible that, at any time, the SIJ may be painless, and moreover, stiff, so that SIJ tests would be expected to be negative in terms of pain provocation.

The gold standard for mobility tests needs to be studied further since the reliability of the measurement methods does not seem to be well established at this time. The use of sensitivity and specificity calculations are a necessary factor when determining whether a study is sound or not. In this review, the authors found only three studies which used these calculations.

Östgaard et al. (1994) was the only study that used prediction values. The use of prediction values, however, has a number of pitfalls. In the other studies, the study designs showed some shortcomings in their construction. In the studies of Fortin and

Table 4. Overview of the validity by author for individual SIJ mobility tests

Test	First Author	Sensitivity	Specificity	Method score	Authors' conclusion
Overtake phenomenon	Egan	na	na	41	Not valid
Gillet test	Dreyfuss	43%	68%	58	Not valid
Long sitting test	Bemis	41%*	83%*	23	Valid

na = not available; *calculated by the authors of this article.

Falco (1997) and Slipman et al. (1998), for example, the research on the study population failed in part because a control group without SIJ dysfunction was not involved. The prevalence of subjects with presumed SIJ dysfunction was high, but predictive positive and negative values could not be counted. The majority of the studies did not show acceptable methodological scores; only two out of 11 studies scored more than 50 points.

The overall negative conclusion of this review can be attributed to the inappropriate design of the studies included. The final results recorded by the authors of the original validity studies for mobility tests and pain provocation tests are disappointing. A precondition for accepting a diagnostic test to be valid and useful in clinical practice is that the test in question must have acceptable reliability. In part 1 of this review, the authors concluded that only the Gaenslen test and the Thigh Thrust test have potential reliability. In the literature, however, there are several reports dealing with validity studies of SIJ diagnostic tests. In our literature research we found only a small number of reports using the selected keywords. Many authors appeared to be confused as to whether their study was designed to test reliability or validity. The authors of this review have, therefore, identified and classified studies into validity or reliability types. There may be some debate about this procedure and the results presented but the authors' final consideration in including all acceptable reports was to diminish publication bias.

Slipman et al. (1998) developed the only multiple-test score (MTS) for validity testing but their design shows some shortcomings. Broadhurst and Bond (1998) suggested that they had carried out MTS but in fact they studied three tests separately and made no attempt to combine these tests. Developing the proper study design for diagnostic test research in the field of SIJ dysfunction is a major challenge for the future. Based on the outcome of part 1 of this review the authors suggest studying provocation tests as a priority.

The choice of SIJ tests validity studies has to be made based on tests which have already been proven reliable. Initially, as stated in part 1, MTS seem to be the most favourable method for inclusion in reliability studies. What is clear is that methodologically sound procedures for carrying out reliability studies are required.

In an ideal situation, investigators should select two groups for validity study designs: a control group which has no SIJ dysfunction and a group of patients that have presumed SIJ dysfunction. The main problem will be in recruiting enough individuals to the SIJ dysfunction group. Pain provocation tests, as well as anaesthetic arthrogram should be applied to both groups. By following this protocol, it will be possible to calculate sensitivity, specificity and

predictive value rates, confidence intervals and likelihood ratios. The prevalence of SIJ dysfunction conditions in relation to a control population in the study design is essential in order to produce meaningful data.

Nowadays, a randomized clinical trial appears to be an inevitable necessity for a diagnostic test that claims a therapeutic impact and may be appropriate for this area of study. Finally, the authors recommend, as stated in part 1, the use of MTS for further research purposes.

CONCLUSION

In this review, only one study (Östgaard et al. 1994) showed a potentially acceptable methodological score: acceptable sensitivity and specificity was found by using the Thigh Thrust test on pregnant women. There is, however no, other data available to support this positive conclusion.

This review of 11 validity studies of SIJ pain provocation and mobility tests showed, in general, the presence of poor methodological quality. The conclusions drawn by the initial authors concerning validity was negative for most tests due to the insufficient reliability of the SIJ tests and to shortcomings in study design. Therefore, at this time, it is questionable whether any SIJ tests are of any value for clinical practice.

Acknowledgements

The authors wish to express their thanks to Su Carleton and Dirk Kokmeyer, for their most helpful assistance in the preparation of this paper.

References

- Alviso DJ, Dong GT, Lentell GL 1988 Intertester reliability for measuring pelvic tilt in standing. *Physical Therapy* 68(9): 1347-1351
- Bemis T, Monte D 1987 Validation of the long sitting test on subjects with iliosacral dysfunction. *Journal of Orthopaedic and Sports Physical Therapy* 8(7): 336-345
- Blower PW, Griffin AJ 1984 Clinical sacroiliac tests in ankylosing spondylitis and other causes of low back pain - 2 studies. *Annales of Rheumatic Disorders* 43: 192-195
- Broadhurst NA, Bond MJ 1998 Pain provocation tests for the assessment of sacroiliac joint dysfunction. *Journal of Spinal Disorders* 11(4): 341-345
- Burdett RG, Brown KE, Fall MP 1986 Reliability and validity of four instruments for measuring lumbar spine and pelvic positions. *Physical Therapy* 66(5): 677-684
- Cibulka MT, Delitto A, Koldenhoff RM 1988 Changes in innominate tilt after manipulation of the sacroiliac joint in patients with low back pain. *Physical Therapy* 68(9): 1359-1363
- Dreyfuss P, Dreyer S, Griffin J, Hoffman J, Walsh N 1994 Positive sacro-iliac screening tests in asymptomatic adults. *Spine* 19(10): 1138-1143
- 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(22): 2594-2602

- Egan D, Cole J, Twomey L 1996 The standing forward flexion test: an inaccurate determinant of sacroiliac joint dysfunction. *Physiotherapy* 82(4): 236-242
- Fortin JD, Falco FJE 1997 The Fortin finger test: an indicator of sacroiliac pain. *American Journal of Orthopedics* 26(7): 477-480
- Gajdosik R, Simpson R, Smith R, DonTigny RL 1985 Pelvic tilt. Intratester reliability of measuring the standing position and range of motion. *Physical Therapy* 65(2): 169-173
- Gemmell HA, Jacobson BH 1990 Incidence of sacroiliac joint dysfunction and low back pain in fit college students. *Journal of Manipulative and Physiological Therapeutics* 13(2): 63-67
- Hungerford BA, Gillear WL 1998 Sacroiliac joint angular rotation during the Stork test and Hip drop test in normal subjects: pilot study results. In: Vleeming et al. (eds) 3rd World Congress on Low Back and Pelvic Pain. Vienna, pp. 332-335
- Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F 1994 Guidelines for meta-analysis evaluating diagnostic tests. *Annals of Internal Medicine* 120(8): 667-676
- Jaeschke R, Guyatt G, Sackett DL 1994a Users guides to the medical literature. III How to use an article about a diagnostic test. A. Are the results of the study valid. *Journal of the American Medical Association* 271(5): 289-391
- Jaeschke R, Guyatt G, Sackett DL 1994b Users guides to the medical literature. III How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients. *Journal of the American Medical Association* 271(9): 703-707
- Koes B 1994 Meetinstrumenten en wetenschappelijk onderzoek. In: Aufdenkampe G et al. (eds) *Jaarboek Fysiotherapie en Kinesitherapie*. Bohn Stafleu Van Loghum, Houten pp 49-58
- Landis RJ, Koch GG 1977 The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174
- Laslett M 1998 Letter to the editor. *Spine* 23(8): 962-963
- Maigne JY, Aivaliklis A, Pfefer F 1996 Results of sacroiliac double block and value of sacroiliac pain provocation tests in 54 patients with low back pain. *Spine* 21(16): 1889-1892
- Mierau DR, Cassidy JD, Hamin T, Milne RA 1984 Sacroiliac joint dysfunction and low back pain in school age children. *Journal of Manipulative Physiological Therapeutics* 7(2): 81-84
- Östgaard HC, Zetherström G, Roos-Hansson E 1994 The posterior pelvic pain provocation test in pregnant women. *European Spine Journal* 3: 258-260
- Rantanen P, Airaksinen JM 1985 Poor agreement between so-called sacroiliac joint tests in ankylosing spondylitis patients. *Journal of Manual Medicine* 4: 62-64
- Reid MC, Lachs MS, Feinstein AR 1995 Methodological standards in diagnostic test research. Getting better but still not good. *Journal American Medical Association* 274(8): 645-651
- Russell AS, Maksymovich W, LeClerq S 1981 Clinical examination of the sacroiliac joints: A prospective study. *Arthritis Rheumatism* 24: 1575-1577
- Slipman CW, Sterenfild EB, Chou LH, Herzog R, Vresilovic E 1998 The predictive value of provocative sacroiliac joint stress maneuvers in the diagnosis of sacroiliac joint syndrome. *Archives of Physical Medicine and Rehabilitation* 79: 288-292
- Sturesson B 1998 Mobility in the sacroiliac joints an approach for understanding. In: Vleeming et al. (eds.) 3rd World Congress on Low Back and Pelvic Pain. Vienna, pp 132-140
- Tanner J 1997 Letter to the editor. *Spine* 22(14): 1673
- van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM and the Editorial Board of the Cochrane Collaboration Back Review Group 1997 *Spine* 22(20): 2323-2330
- van der Wurff P, Meyne W, Hagmeyer RHM 1999 Clinical tests of the sacroiliac joint. A systematic methodological review. Part I Reliability. *Manual Therapy* 5(1): 30-36