Systematic review of patient history and physical examination to diagnose chronic low back pain originating from the facet joints

Maas, E. T.; Juch, J.N.S.; Ostelo, R. W. J. G.; Groeneweg, J.G.; Kallewaard, J. W.; Koes, Bart W; Verhagen, Arianne P; Huygen, F. J. P. M.; van Tulder, M. W.

published in
European Journal of Pain
2017

DOI (link to publisher)
10.1002/ejp.963

document version
Publisher's PDF, also known as Version of record

EJP
European Journal of Pain

REVIEW ARTICLE

Systematic review of patient history and physical examination to diagnose chronic low back pain originating from the facet joints

E.T. Maas[1], J.N.S. Juch[2], R.W.J.G. Ostelo[1,3], J.G. Groeneweg[2], J.W. Kallewaard[4], B.W. Koes[5], A.P. Verhagen[5], F.J.P.M. Huygen[2], M.W. van Tulder[1]

1 Department of Health Sciences and the EMGO+ Institute for Health and Care Research, Faculty of Earth and Life Sciences, VU University Amsterdam, The Netherlands
2 Department of Anaesthesiology, Erasmus Medical Centre, Rotterdam, The Netherlands
3 Department of Epidemiology and Biostatistics, VU University Medical Centre Amsterdam, The Netherlands
4 Rijnstate Hospital, Velp, The Netherlands
5 Department of General Practice, Erasmus Medical Centre, Rotterdam, The Netherlands

Correspondence
Esther T. Maas
E-mail: esther.maas@vu.nl

Funding sources
The Netherlands Organisation for Health Research and Development (ZonMw) (Nm: 171202013).

Conflicts of interest
None declared.

Accepted for publication
3 September 2016

doi:10.1002/ejp.963

Abstract

Patient history and physical examination are frequently used procedures to diagnose chronic low back pain (CLBP) originating from the facet joints, although the diagnostic accuracy is controversial. The aim of this systematic review is to determine the diagnostic accuracy of patient history and/or physical examination to identify CLBP originating from the facet joints using diagnostic blocks as reference standard. We searched MEDLINE, EMBASE, CINAHL, Web of Science and the Cochrane Collaboration database from inception until June 2016. Two review authors independently selected studies for inclusion, extracted data and assessed the risk of bias. We calculated sensitivity and specificity values, with 95% confidence intervals (95% CI). Twelve studies were included, in which 129 combinations of index tests and reference standards were presented. Most of these index tests have only been evaluated in single studies with a high risk of bias. Four studies evaluated the diagnostic accuracy of the Revel's criteria combination. Because of the clinical heterogeneity, results were not pooled. The published sensitivities ranged from 0.11 (95% CI 0.02–0.29) to 1.00 (95% CI 0.75–1.00), and the specificities ranged from 0.66 (95% CI 0.46–0.82) to 0.91 (95% CI 0.83–0.96). Due to clinical heterogeneity, the evidence for the diagnostic accuracy of patient history and/or physical examination to identify facet joint pain is inconclusive. Patient history and physical examination cannot be used to limit the need of a diagnostic block. The validity of the diagnostic facet joint block should be studied, and high quality studies are required to confirm the results of single studies.

Significance: Patient history and physical examination cannot be used to limit the need of a diagnostic block. The validity of the diagnostic facet joint block should be studied, and high quality studies are required to confirm the results of single studies.

1. Introduction

Chronic low back pain (CLBP) is a widespread problem with major social and economic consequences (Martin et al., 2008; Manchikanti et al., 2009). Over the last decades multiple structures in the lumbar spine have been considered to cause CLBP. Goldthwait et al. first described the facet joints as the source of pain in 1911, after which Ghormley introduced the term 'facet syndrome' in 1933 (Goldthwait, 1911; Ghormley, 1933). Facet joint pain is defined as pain resulting from any integral structure of the facet joints, including the fibrous capsule, synovial membrane, hyaline cartilage surfaces and bony articulation (Cohen and Raja, 2007). The prevalence of facet joint pain is estimated at 15–41% of the CLBP population, and varies widely in the literature depending on setting, definition and diagnostic procedures (Schwarzer et al., 1994a; Manchikanti et al., 1999, 2004; Eubanks et al., 2007; Hicks et al., 2009; DePalma et al., 2011).

History taking and physical examination are two commonly used procedures in diagnosing facet joint pain and are considered the index test in this review. The validity and reliability of history taking and physical examination in clinical practice are controversial (Cohen and Raja, 2007; Hancock et al., 2007). Physicians and therapists use the information gained from history taking and physical examination to decide on the use of further diagnostic tests. Imaging (plain radiography, computed tomographic or magnetic resonance imaging), and diagnostic blocks are tests subsequent to history taking and physical examination, and are used in clinical practice to diagnose, amongst others, facet joint pain. The evidence of radiological imaging to predict response to diagnostic facet joint blocks and diagnose CLBP originating from the facet joints has been shown to be conflicting at best, and are therefore not used in this review (Carrera, 1980; Fairbank et al., 1981; Carrera and Williams, 1984; Raymond et al., 1984; Lewinnek and Warfield, 1986; Helbig and Lee, 1988; Jackson et al., 1988; Murtagh, 1988; Revel et al., 1992; Schwarzer et al., 1995; Dolan et al., 1996; Itz et al., 2016).

The most commonly used test to diagnose CLBP originating from the facet joints is the diagnostic block. The rationale for these blocks is to anaesthetize a painful facet joint for the duration of the anaesthetic effect (Bogduk, 1983; Manchikanti et al., 2000b). The diagnostic block is recommended in many guidelines, among others in the Dutch guideline for anaesthesiology (Itz et al., 2016). Although not a gold standard, diagnostic blocks are currently the best tests available for diagnosing facet joint pan, and the reference standard in this review (Schwarzer et al., 1994b; Hogan and Abram, 1997; Manchikanti et al., 2000b; Dreyfuss et al., 2003; Cohen et al., 2012).

The last systematic review that summarized the diagnostic accuracy of patient history and/or physical examination was published more than 8 years ago. It is important to update this review, because including more recent publications may have an impact on the overall evidence. Furthermore, patient care can be improved if patient history and physical examination could limit the need for further invasive diagnostic procedures and treatments. Also, the methods of diagnostic systematic reviews evolved and this update will use up-to-date methodology. This review will provide an overview of the current practice in diagnosing CLBP originating from the facet joints.

The objective of this review is to assess the diagnostic accuracy of patient history and/or physical examination, compared to a diagnostic block, to identify chronic low back pain originating from the facet joints.

2. Literature search methods

2.1 Design

A systematic review of diagnostic accuracy studies.

2.2 Data sources and search strategy

We performed a database search using Medline (in OvidSP), EMBASE, CINAHL, Web of Science, Google Scholar and the Cochrane Collaboration database from their date of inception until June 30th 2016 using MeSH terms (Medline), thesaurus (EMBASE, CINAHL, Web of Science) and free-text words (Cochrane, Google Scholar). Search terms were related to the diagnostic accuracy, facet joint pain, index and reference tests (for details, see Supporting Information Appendix S1). No method filter was included, because using filters might lead to missing relevant studies (Beynon et al., 2013). Based on the results of the electronic search, the bibliographies of all systematic reviews and eligible diagnostic studies were reviewed. Two experts in the field of diagnostic testing in anaesthesiology and LBP evaluated the study selection independently of each other to ensure that the search was comprehensive.

2.3 Study selection

Two review authors (JJ and EM) independently screened the search results based on title, key words

and abstract. We obtained full texts for hits that were considered relevant by at least one of the authors. Disagreements were resolved by consensus. In case of persistent disagreement or uncertainty, a third reviewer (AV) made the final decision. Reasons for exclusion were noted. The studies had to fulfil the following criteria for final inclusion in the review:

(1) Study design: retrospective, prospective or cross-sectional studies.
(2) Data collection: studies designed using existing data as well as studies designed specifically to address the diagnostic question with newly collected data were included. No restrictions with regard to year of publication or language were applied. Non-English language reports were categorized as 'awaiting assessment' until appropriate translation was obtained. Papers published only in abstract form, case reports, as well as animal and post-mortem studies were excluded.
(3) Study population: adult patients, of either gender, suffering from CLBP were included. Patients with acute trauma, fractures, malignancies and inflammatory diseases were excluded.
(4) Index test: all history taking and physical examination tests in the included studies were analysed, in isolation and in combination.
(5) Reference standard: A diagnostic block of the medial branch of the dorsal ramus under X-ray or low volume intra-articular blocks, as recommended by the Dutch guideline for anaesthesiology (Itz et al., 2016). There is variation in the use of diagnostic blocks (Cohen and Raja, 2007; Cohen et al., 2012). In general, pain reduction in 50% or more implies that the anaesthetized joint is the source of the pain, but studies using other thresholds were included as well. Single and double diagnostic blocks (to confirm the results of the first test) have comparable validity and were therefore both included as reference standard (van Wijk et al., 2005; Itz et al., 2016).

Information on the inter- or intra-observer reproducibility of the tests, or both, if reported or referenced in the study were collected. The maximum time interval between the index test and reference standard was 3 months, as no change in symptoms in CLBP patients was expected in this time range.

2.4 Data extraction

Two reviewers (JJ and EM) developed and completed the data extraction form.

Data were extracted on:

(1) Author, date of publication, location, journal
(2) Study design
(3) Study population characteristics: basic demographics, number of patients (number eligible for the study, number enrolled in the study, number receiving index test and reference standard, number of whom the results are reported in the two-by-two table, reasons for withdrawal), inclusion and exclusion criteria, setting
(4) Index tests and reference standard characteristics: type of test, method of execution, cut-off-value (outcome scales), positivity thresholds
(5) Outcomes: true-positives, false-positives, true-negatives and false-negatives

2.5 Quality assessment

Two reviewers (JJ and EM) independently assessed the methodological quality using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (Whiting et al., 2003, 2011). Signalling questions were used to assess the risk of bias in four domains: patient selection (three signalling questions), index tests (two signalling questions), reference standard (two signalling questions), and flow and timing (four signalling questions). We added one extra item to the QUADAS-2 tool to assess reliability: 'Were data on observer variation reported and within an acceptable range?'. When there was at least one 'no' or 'unclear' response to a signalling question for a given domain, we scored the risk of bias domain as high or unclear, respectively. Furthermore, concerns about applicability of the studies were scored on three domains: patient selection, index test and reference standard.

Disagreements were resolved by consensus or by an independent third author (RO). No summary scores were used (Jüni et al., 1999; Whiting et al., 2005). All domains covered by the QUADAS-2 tool were considered as potential covariates. That is, if an item was assumed to potentially influence the reported sensitivity and specificity, this item was incorporated in the bivariate analyses to examine the effect of this potential source of bias on the diagnostic accuracy of patient history and physical examination. The items of the QUADAS-2 tool are displayed in Supporting Information Appendix S2.

2.6 Data synthesis and analysis

Study-specific estimates of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), as well as sensitivity and specificity for each index test with 95% confidence intervals (CI)

were calculated. In case of clinical homogeneity (same index tests and reference standards, and a comparable study population) a multivariate meta-analysis was conducted using a hierarchical summary receiver operating characteristic (HSROC) model. Pooled estimates of sensitivity and specificity and 95% CIs were calculated with this model (Reitsma et al., 2005). The HSROC model is recommended for analysing sensitivity and specificity data reporting on more than one threshold for positive tests (Dukic and Gatsonis, 2003). The HSROC model estimates the between-study variability (heterogeneity), allowing for the theoretically expected non-independence of sensitivity and specificity across studies, and can therefore be seen as an improvement of the simple summary receiver operating characteristic approach (Dukic and Gatsonis, 2003; Reitsma et al., 2005).

For each index test, parameter estimates from the fitted model were inputted into the software program Revman5 to graph the ROC space. We investigated the potential influence of differences in study population, index test and reference standard criteria, differences in time interval of the evaluation of pain reduction, and study design features (prospective vs. retrospective).

When studies were clinically heterogeneous, no meta-analysis was performed, but only the range of sensitivity and specificity and the 95% CI (as published) was presented. Furthermore, the authors assessed if differences in diagnostic accuracy of patient history and/or physical examination could be explained by variation in reference standard procedures.

3. Results

3.1 Study selection

The study flow diagram is presented in Fig. 1. Database searching yielded 867 individual papers and four additional papers were identified through reference checking. After removal of duplicates and exclusions based on title and abstract, 26 full-text papers were assessed for eligibility. Fourteen papers were excluded for various reasons (reasons are shown in Fig. 1), which resulted in 12 included papers.

Two studies reported on the same study population (Schwarzer et al., 1994a, c), and one study (Laslett et al., 2006) was part of a larger diagnostic accuracy study (Laslett et al., 2005).

3.2 Description of the studies

Details on the population, setting, index test, reference standard and target condition are presented in Supporting Information Appendix S3. The 12 included studies consisted of ten cross-sectional studies, one case–control study (González, 2004), and one retrospective cohort (Jung et al., 2007), in which information was collected from medical records. The sample size of the studies ranged from 23 (Young et al., 2003) to 259 (Jung et al., 2007) patients that we could use in this review, and the mean age ranged from 38 (Schwarzer et al., 1994c) to 62 (Jung et al., 2007) years of age. The index test and reference standard were performed on the same occasion or the timing was unclear. The maximum timing between both diagnostic blocks was 2 weeks.

A variety of patient history and physical examination items was used as index test. Index tests found in more than one study were: non-centralization, onset of trauma, age>65 years, pain well relieved by recumbency, pain not exacerbated by coughing, pain not exacerbated by forward flexion, pain not exacerbated by extension, pain not exacerbated by rising from flexion, and pain not exacerbated with the extension-rotation test. Five studies reported findings on diagnostic accuracy on combinations of test results. Four studies reported on Revel's criteria (Revel et al., 1992, 1998; Manchikanti et al., 2000a; Laslett et al., 2004). Index tests that were evaluated only in single studies include: a pain distribution pattern (Jung et al., 2007), a clinical prediction rule (Laslett et al., 2005), the new lumbar facet sign (González, 2004), and many aspects of a physical examination. All index tests compared to the reference standards are shown in Supporting Information Appendix S4.

Seven studies used a single diagnostic block as reference standard. Four studies used a second confirmatory diagnostic block in patients with a positive first diagnostic block (Schwarzer et al., 1994a; Manchikanti et al., 1999, 2000a; González, 2004). One study compared pain distribution patterns to a single diagnostic block, a double diagnostic block and an overall diagnostic rate (Jung et al., 2007). The cut-off point for a positive diagnostic block ranged from 'clinical improvement' (González, 2004) to 'more than 95% pain reduction 30–45 min after procedure' (Laslett et al., 2006).

3.3 Methodological quality

The results of the quality assessment are presented in Fig. 2. The two reviewers (JJ and EM) agreed on

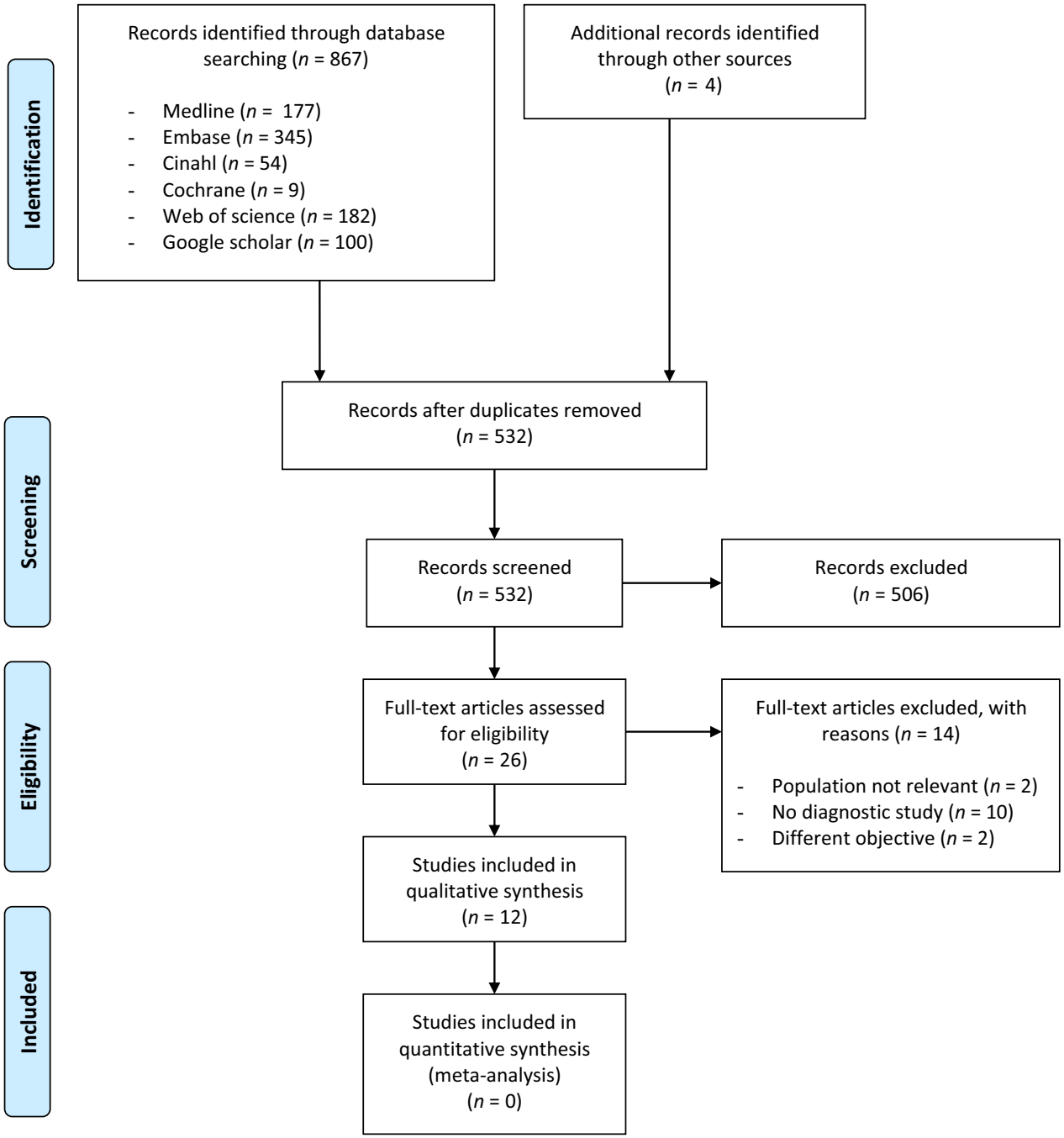

Figure 1 Study flow diagram.

39 of the 48 risk of bias domains (80%) and on all applicability concern domains while scoring the QUADAS-2 tool.

3.3.1 Risk of bias

We scored patient selection 'unclear' or 'high risk of bias' in eight of the 12 studies. In four studies (Revel et al., 1992, 1998; Schwarzer et al., 1994a, c), the patient selection was adequately described and scored 'low risk of bias'. Half of the studies scored 'low risk of bias' on description of the index test. Only the study of González (2004) scored 'high risk of bias' on this domain as in this study the index test was interpreted with prior knowledge of the

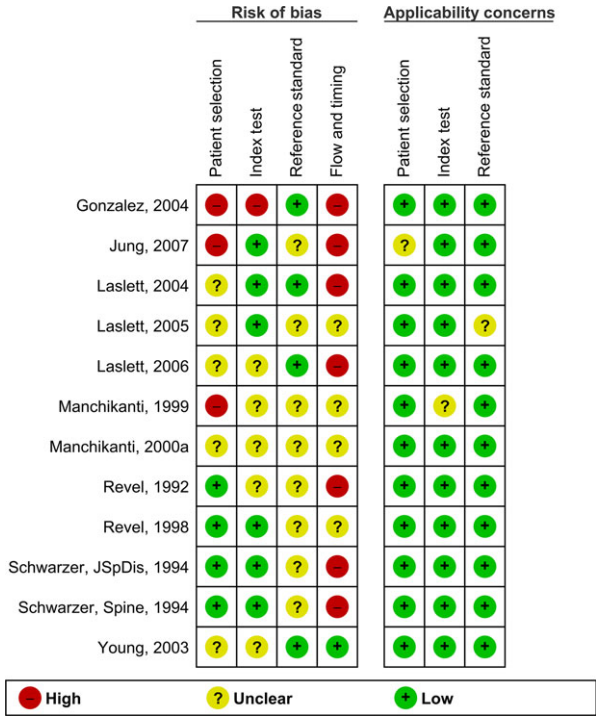

Figure 2 Quality assessment summary by the QUADAS-2 tool.

reference standard results. It remained unclear if a pre-specified threshold was used for the scoring of the index test in four studies, resulting in an 'unclear risk of bias' score. In eight studies it was unclear if the reference standard results were interpreted without knowledge of the index test results. We scored seven studies as 'high risk of bias' on the flow and timing domain. This was mostly due to the fact that not all patients received a reference standard or received the same reference standard.

3.3.2 Applicability concerns

For the assessment of applicability, there was no concern for nine studies that the included patients, setting, the conduct and interpretation of the index test, and the target condition (as defined by the reference standard) in each of the included studies did not match the review question. In one study (Jung et al., 2007) the patient selection was insufficiently described, in another study (Laslett et al., 2005) the reference standard was unclear, and in a third study (Manchikanti et al., 1999) the description of the index test was unclear.

3.4 Findings

In Supporting Information Appendix S4 shows the index tests, reference standards, sample size, true positives, true negatives, false positives, false negatives, prior probability, sensitivity and specificity for all studies in which it was possible to create a 2 × 2 table. Ten index tests were evaluated in more than one study and are shown in Table 1. Because of clinical heterogeneity, a meta-analysis was not considered relevant.

3.5 Diagnostic value of index tests for facet joint pain described in two or more studies

3.5.1 Non-centralization

The non-centralization phenomenon (Sufka et al., 1998; Werneke and Hart, 2001) was evaluated as index test in two studies (Young et al., 2003; Laslett et al., 2006). Both studies used a single diagnostic block as reference standard, while Laslett et al. (2006) used a cut-off point of 80% pain reduction and Young et al. (2003) of 95% pain reduction. Both studies did not have false negative cases, which resulted in a sensitivity of 1.00. Specificity was poor in both studies: 0.17 (95% CI 0.11–0.27) in Laslett et al. (2006); 0.11 (95% CI 0.02–0.44) in Young et al. (2003).

3.5.2 Traumatic onset

Trauma as a cause for facet joint pain was evaluated in two studies (Manchikanti et al., 1999, 2000a). The index test was compared to a controlled double diagnostic block in both studies. Sensitivity and specificity were poor with a maximum sensitivity of 0.54 (95% CI 0.41–0.66) and maximum specificity of 0.50 (95% CI 0.41–0.59).

3.5.3 Revel's criteria separately

Revel's criteria include (1) age over 65 years, (2) pain well relieved by recumbency and pain not exacerbated by (3) coughing, (4) forward flexion, (5) extension, (6) rising from flexion, (7) extension-rotation (Revel et al., 1998). All tests were studied by Revel (Revel et al., 1992, 1998) and/or Manchikanti et al. (1999, 2000a). Because of the heterogeneity in the reference standard (the use of single and double diagnostic blocks), no pooled results are presented. Sensitivity ranged from 0.15 (95% CI 0.09–0.25) to 1.00 (95% CI 0.77–1.00); specificity ranged from 0.13 (95% CI 0.08–0.20) to 0.86 (95% CI 0.79–0.91).

3.5.4 Revel's criteria combined

Four studies evaluated the performance of the combined Revel's criteria (positive with five or more

Table 1 Diagnostic value of index tests for facet joint pain in two or more studies.

Study	Reference standard	Sample size	TP	TN	FP	FN	Prior probability	Sensitivity (95%CI)	Specificity (95%CI)
	Non- centralization								
Laslett, 2006	≥95% pain reduction after diagnostic block	92	11	14	67	0	0.12	1.00 (0.74–1.00)	0.17 (0.11–0.27)
Young, 2003	≥80% pain reduction after diagnostic block	23	14	1	8	0	0.61	1.00 (0.78–1.00)	0.11 (0.02–0.44)
	Traumatic onset								
Manchikanti, 1999	≥75% pain reduction after double diagnostic block	120	29	31	35	25	0.45	0.54 (0.41–0.66)	0.47 (0.35–0.59)
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	200	40	58	58	44	0.42	0.48 (0.37–0.58)	0.50 (0.41–0.59)
	Age >65 years								
Manchikanti, 1999	≥75% pain reduction after double diagnostic block	120	10	54	12	44	0.45	0.19 (0.10–0.31)	0.82 (0.71–0.89)
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	200	18	99	17	66	0.42	0.21 (0.14–0.31)	0.85 (0.78–0.91)
Revel, 1998	≥75% pain reduction after lidocaine block	42	5	23	6	8	0.31	0.38 (0.18–0.64)	0.74 (0.62–0.90)
	Pain well relieved by recumbency								
Revel, 1992	≥75% pain reduction after lidocaine block	40	20	8	10	2	0.55	0.91 (0.72–0.97)	0.44 (0.25–0.66)
Revel, 1998	≥75% pain reduction after lidocaine block	42	12	7	22	1	0.31	0.92 (0.67–0.99)	0.24 (0.12–0.42)
	Pain not exacerbated by coughing								
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	200	76	15	101	8	0.42	0.90 (0.82–0.95)	0.13 (0.08–0.20)
Revel, 1992	≥75% pain reduction after lidocaine block	40	18	9	9	4	0.55	0.82 (0.62–0.93)	0.50 (0.29–0.71)
Revel, 1998	≥75 pain reduction after lidocaine block	42	13	10	19	0	0.31	1.00 (0.77–1.00)	0.35 (0.20–0.53)
	Pain not exacerbated by forward flexion								
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	200	13	95	21	71	0.42	0.15 (0.09–0.25)	0.82 (0.74–0.88)
Revel, 1992	≥75% pain reduction after lidocaine block	40	14	14	4	8	0.55	0.64 (0.41–0.83)	0.78 (0.52–0.94)
Revel, 1998	≥75% pain reduction after lidocaine block	42	13	14	15	0	0.31	1.00 (0.77–1.00)	0.48 (0.31–0.66)
	Pain not exacerbated by extension								
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	200	8	100	16	76	0.42	0.10 (0.05–0.18)	0.86 (0.79–0.91)
Revel, 1992	≥75% pain reduction after lidocaine block	40	12	13	5	10	0.55	0.55 (0.35–0.73)	0.72 (0.49–0.88)
Revel, 1998	≥75% pain reduction after lidocaine block	42	12	18	11	1	0.31	0.92 (0.67–0.98)	0.62 (0.44–0.77)
	Pain not exacerbated by rising from flexion								
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	210	48	59	63	40	0.42	0.55 (0.44–0.65)	0.48 (0.39–0.57)
Revel, 1992	≥75% pain reduction after lidocaine block	40	17	10	8	5	0.55	0.77 (0.57–0.90)	0.56 (0.34–0.75)
Revel, 1998	≥75% pain reduction after lidocaine block	42	13	17	12	0	0.31	1.00 (0.77–1.00)	0.59 (0.41–0.74)
	Pain not exacerbated with the extension-rotation test								
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	200	57	35	81	27	0.42	0.68 (0.57–0.77)	0.30 (0.23–0.39)

Table 1 (*Continued*)

Study	Reference standard	Sample size	TP	TN	FP	FN	Prior probability	Sensitivity (95%CI)	Specificity (95%CI)
Revel, 1998	≥75% pain reduction after lidocaine block Revel's criteria combined[a]	42	10	14	15	3	0.31	0.76 (0.50–0.92)	0.48 (0.31–0.66)
Laslett, 2004	75% pain reduction after diagnostic block	116	3	81	8	24	0.23	0.11 (0.02–0.29)	0.91 (0.83–0.96)
Manchikanti, 2000a	≥75% pain reduction after double diagnostic block	200	11	97	18	74	0.43	0.13 (0.07–0.22)	0.84 (0.76–0.90)
Revel, 1992	≥75% pain reduction after Lidocaine block	40	14	16	2	8	0.55	0.64 (0.41–0.83)	0.89 (0.65–0.99)
Revel, 1998	≥75% pain reduction after Lidocaine block	42	13	19	10	0	0.31	1.00 (0.75–1.00)	0.66 (0.46–0.82)

TP, true positives; FP, false positives; TN, true negatives; FN, false negatives.

[a]Revel's criteria include (1) age over 65 years, (2) pain well relieved by recumbency and pain not exacerbated by (3) coughing, (4) forward flexion, (5) extension, (6) rising from flexion, (7) extension-rotation. Revel's criteria are positive with five or more clinical characteristics.

clinical characteristics) (Revel et al., 1992, 1998; Manchikanti et al., 2000a; Laslett et al., 2004). Because of the clinical heterogeneity (variation in the reference standard), results were not pooled. Sensitivity ranges from 0.11 (95% CI 0.02–0.29) to 1.00 (95% CI 0.75–1.00), and the specificity ranges from 0.66 (95% CI 0.46–0.82) to 0.91 (95% CI 0.83–0.96) (Fig. 3).

3.5.5 Diagnostic value of index tests for facet joint pain presented in one study only

For 48 of the 129 index test and reference standard combinations that were described in only one study, it was impossible to construct a 2 × 2 table. Results of the diagnostic accuracy of all index tests that were described in one study only are shown in Supporting Information Appendix S4.

In the study of Laslett et al. (2006), the extension-rotation test was compared to five thresholds of reference standards: 75%, 80%, 85%, 90% and 95% pain reduction after a diagnostic block. Almost no change in specificity was shown due to the very little change in true negatives. Neither was a difference in sensitivity shown (Fig. 4).

4. Discussion

This review summarizes 12 diagnostic accuracy studies, including 129 combinations of patient history and/or physical examination and diagnostics blocks, to identify the facet joints as the source of CLBP. Most index tests have only been evaluated in small single studies with a high risk of bias and provide insufficient evidence about the diagnostic accuracy of these tests. The diagnostic accuracy of Revel's criteria was evaluated in four studies, which only concerned two research teams. Sensitivity ranged from 0.15 (95% CI 0.09–0.25) to 1.00 (95% CI 0.77–1.00); specificity ranged from 0.13 (95% CI 0.08–0.20) to 0.86 (95% CI 0.79–0.91). No meta-analyses were performed due to clinical heterogeneity.

4.1 Factors affecting interpretation

Firstly, one reason for clinical heterogeneity in this review could be the wide range in prior probabilities of having CLBP originating from the facet joints in the included studies (11.9–55.0%) (Revel et al., 1992; Laslett et al., 2006). Sensitivity and specificity can vary between studies with different prior

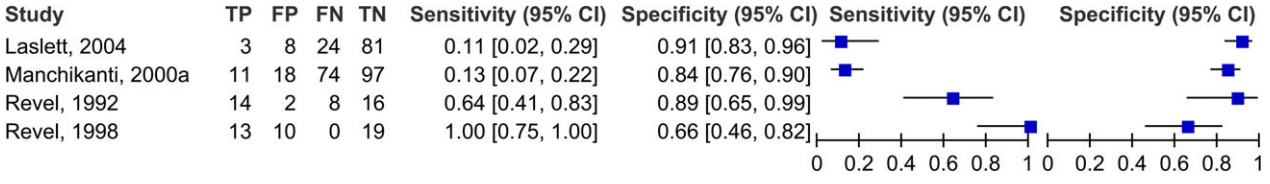

Figure 3 Revel's criteria – reference test (single diagnostic block in the studies of Laslett et al. (2004), Revel et al. (1992) and Revel et al. (1998). Double diagnostic block in the study of Manchikanti et al. (2000a)).

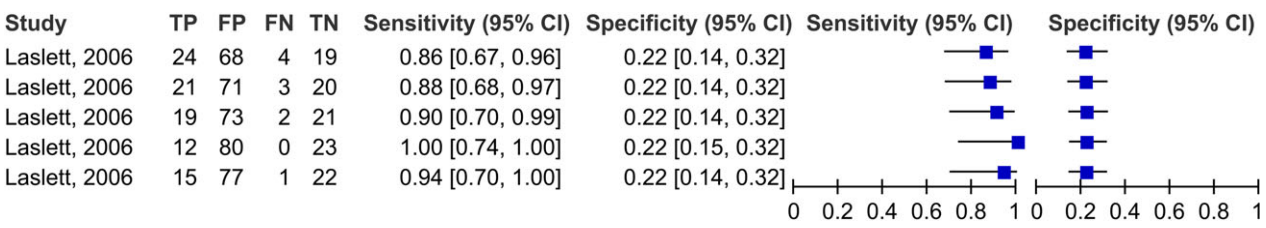

Figure 4 Extension-rotation test – reference test: 1. 75%, 2. 80%, 3. 85%, 4. 90% and 5. 95% pain reduction after a diagnostic block (Study by Laslett et al. (2006) (N = 120)).

probabilities of disease (Leeflang et al., 2013). Furthermore, the study setting and patient selection was only described clearly in four of 12 included studies. This could have affected the results as well, as patients in the other eight included studies may not have been a representative sample. Moreover, it is important to state that the results of this review can only be interpreted in a secondary care context. Patient history and physical examination are conducted in multiple settings, but diagnostic blocks are used to predict the response on radiofrequency denervation. Radiofrequency denervation is mostly performed at the pain clinic, which is a secondary care setting.

Secondly, recent studies included in this review showed results on many index tests that were only evaluated in single studies. Very few studies were reproduced to confirm these results. The results of this review should therefore be interpreted with caution. Additionally, the reliability of the index tests and the (inter- and intra-) observer variation may have affected the results due to a lack of consensus in procedures and thresholds. Risk of bias was considered high on the 'index test' item in six studies because of the lack of positive or negative test result definition.

Thirdly, there is no gold standard for diagnosing facet joint pain. In this review the diagnostic facet joint block was used as reference test. Although little evidence is available that truly justifies the use of diagnostic blocks as the gold standard, it is currently the best tool available. We found variation in the execution of the reference standard. Therefore, we assessed the heterogeneity in diagnostic blocks, and if this would explain the differences in sensitivity and specificity for the index test. We could not conclude that this was the case. Still it is important to stress that the lack of a gold standard represents a flaw in diagnostic studies. As the index test can never perform better than the reference standard, its value may be underestimated (Riegelman, 2005). Despite the lack of validity and the chance of false-positive tests (Schwarzer et al., 1994b; Cohen et al., 2012), diagnostic blocks are frequently performed in daily practice and currently the best tests available for diagnosing CLBP originating from the facet joints (Itz et al., 2016).

The results of this review are in line with the diagnostic accuracy of patient history and/or physical examination for identifying other sources of CLBP, such as disc herniation or sacroiliac joint pain (Hancock et al., 2007; Van der Windt et al., 2010). The diagnostic value of physical examination tests in primary care populations and other general, unselected patient groups is still unclear as well, as evidence from these settings is scarce.

4.2 Reliability

This review focused on the diagnostic performance (i.e. validity) of patient history and/or physical examination in patients with CLBP. The reliability of these tests was outside the scope of this review. However, adequate reliability (inter- and intra-observer agreement) is a prerequisite for good performance of diagnostic tests.

Only three studies provided some information on reliability (Schwarzer et al., 1994a, c; Laslett et al., 2005). Laslett et al. (2005) reported weak inter-observer agreement for the physiotherapist diagnosis of facet joint pain with a kappa value of 0.31 (95% CI: 0.18–0.44). Both studies of Schwarzer et al. (1994a, b, c) reported substantial to excellent inter-observer agreement in physical examination between the principal investigator and the other investigators. The κ scores for comparisons between observers ranged between 0.69 and 1.00.

4.3 Strengths and limitations

A strength of this review is the assessment of methodological quality by two independent assessors using the QUADAS-2 tool as state-of-the-art method. Careful assessment of the quality of included studies is essential in a systematic review and increases the validity of the review. Secondly, we used a comprehensive search strategy. The use of search filters was avoided; which minimized the risk of missing relevant studies.

The most important limitation of this review was the relatively small number of clinical homogeneous studies. Since the last systematic review on this topic in 2007 (Hancock et al., 2007), six new studies have been published and were included in this review. This doubled the number of studies regarding facet joint pain. However, clinical heterogeneity among the included studies was substantial. For this reason, only the results of the Revel's criteria could be pooled. All other index tests were evaluated in single studies, and we were not able to conclude on the diagnostic accuracy of these index tests. Consequently, the improvement of evidence on the diagnostic accuracy of patient history and physical activity in diagnosing facet joint pain compared to the previous review in 2007 is limited. A second limitation of this review is the poor description of the physical examination in most studies. It often remained unclear what thresholds were used to define positive test results. This may have led to different positive and negative test results between studies.

4.4 Implications for practice

The importance of the findings should be interpreted in the light of its clinical consequences. To cite Revel et al., 1998 (Revel et al., 1998): A set of five clinical characteristics could be used in future randomized controlled studies as a criterion for selecting patients with probable painful facet joints that will be well-relieved by facet joint anaesthesia. However, these clinical characteristics should not be considered as diagnostic criteria for low back pain of facet joint origin in clinical practice with individual patients'. According to the results in this review, not much has changed in the past 17 years. There still is a great variety in the performance of patient history and physical examination. Based on this review it can be concluded that it is unclear what drives the decision to perform a diagnostic block. Patient history and physical examination can only give a modest direction to the facet joint pain diagnosis and not limit the use of the more invasive diagnostic facet joint block.

4.5 Implications for research

New studies on this topic have not led to more clarity, only to more heterogeneity. The evidence for the diagnostic accuracy of patient history and/or physical examination identified by this review is still inconclusive. To provide more profound evidence on the role of patient history and physical examination, there is a strong need for good quality and accurately reported prospective cohort studies. These studies should especially focus on investigating diagnostic accuracy of existing, commonly used tests, rather that introducing new tests. Implementation of the STARD guidelines will improve reporting of diagnostic studies in future research (Bossuyt et al., 2003; Smidt et al., 2006).

5. Conclusions

The evidence by this review for the diagnostic accuracy of patient history and/or physical examination to identify facet joint pain is inconclusive. Studies investigating Revel's criteria found conflicting evidence. Other elements of patient history and physical examination have been investigated in single studies with mostly a high risk of bias. This review provides important recommendations for practice and research: (1) Due to practice variation, patient history and physical examination can only give a modest direction to the facet joint pain diagnosis and not limit the use of a diagnostic block. (2) Future research should assess the validity of the diagnostic facet joint block to serve as a better reference standard. (3) High quality studies are required to confirm the results of single studies, and (4) accurate reporting of diagnostic studies based on the STARD guideline is essential.

Acknowledgements

We thank Wichor Bramer for his help in designing the search strategy.

Author contributions

All authors contributed to the paper and discussed the results and commented on the manuscript. ET Maas participated in the collection, extraction and analyses of the data, assessment of the methodological quality, discussion of core ideas and writing of the paper. JNS Juch participated in the collection, extraction of the data, assessment of the methodological quality, discussion of core ideas and writing of the paper. RWJG Ostelo participated in the extraction and analyses of the data, discussion of core ideas and writing of the paper. AP Verhagen participated in the extraction and analyses of the data, assessment of the methodological quality, discussion of core ideas and writing of the paper. JG Groeneweg, JW Kallewaard, B Koes, FJPM Huygen and MW van Tulder participated in the discussion of core ideas and writing of the paper.

References

Beynon, R., Leeflang, M.M., McDonald, S., Eisinga, A., Mitchell, R.L., Whiting, P., Glanville, J.M. (2013). Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database Syst Rev* 9, Art. No.: MR000022. DOI: 10.1002/14651858.MR000022.pub3.

Bogduk, N. (1983). The innervation of the lumbar spine. *Spine* 8, 286–293.

Bossuyt, P., Reitsma, J., Bruns, D., Gatsonis, C., Glasziou, P., Irwig, L., Moher, D., Rennie, D., De Vet, H., Lijmer, J. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Ann Intern Med* 138, 1–12.

Carrera, G. (1980). Lumbar facet joint injection in low back pain and sciatica: Preliminary results. *Radiology* 137, 665–667.

Carrera, G., Williams, A. (1984). Current concepts in evaluation of the lumbar facet joints. *Crit Rev Diagn Imaging* 21, 85–104.

Cohen, S., Raja, S. (2007). Pathogenesis, diagnosis, and treatment of lumbar zygapophysial (facet) joint pain. *Anesthesiology* 106, 591–614.

Cohen, S., Huang, J., Brummett, C. (2012). Facet joint pain—Advances in patient selection and treatment. *Nat Rev Rheumatol* 9, 101–116.

DePalma, M., Ketchum, J., Saullo, T. (2011). What is the source of chronic low back pain and does age play a role? *Pain Med* 12, 224–233.

Dolan, A., Ryan, P., Arden, N., Stratton, R., Wedley, J., Hamann, W., Fogelman, I., Gibson, T. (1996). The value of SPECT scans in identifying back pain likely to benefit from facet joint injection. *Rheumatology* 35, 1269–1273.

Dreyfuss, P., Dreyer, S., Vaccaro, A. (2003). Lumbar zygapophysial (facet) joint injections. *Spine J* 3, 50–59.

Dukic, V., Gatsonis, C. (2003). Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics* 59, 936–946.

Eubanks, J., Lee, M., Cassinelli, E., Ahn, N. (2007). Prevalence of lumbar facet arthrosis and its relationship to age, sex, and race: An anatomic study of cadaveric specimens. *Spine* 32, 2058–2062.

Fairbank, J., Park, W., McCall, I., O'Brien, J. (1981). Apophyseal injection of local anesthetic as a diagnostic aid in primary low-back pain syndromes. *Spine* 6, 598–605.

Ghormley, R. (1933). Low back pain with special reference to the articular facets, with presentation of an operative procedure. *JAMA* 101, 1773–1777.

Goldthwait, J. (1911). The lumbo-sacral articulation; an explanation of many cases of" lumbago","sciatica" and paraplegia. *Boston Med Surg J* 164, 365–372.

González, J. (2004). Síndrome facetario lumbar. Nuevo signo de diagnóstico clínico. *Rehabilitación* 38, 168–174.

Hancock, M., Maher, C., Latimer, J., Spindler, M., McAuley, J., Laslett, M., Bogduk, N. (2007). Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. *Eur Spine J* 16, 1539–1550.

Helbig, T., Lee, C. (1988). The lumbar facet syndrome. *Spine* 13, 61.

Hicks, G., Morone, N., Weiner, D. (2009). Degenerative lumbar disc and facet disease in older adults: Prevalence and clinical correlates. *Spine* 34, 1301–1306.

Hogan, Q., Abram, S. (1997). Neural blockade for diagnosis and prognosis: A review. *Anesthesiology* 86, 216–241.

Itz, C., Willems, P., Zeilstra, D., Huygen, F. (2016). Dutch multidisciplinary guideline for invasive treatment of pain syndromes of the lumbosacral spine. *Pain Pract* 16, 90–110.

Jackson, R., Jacobs, R., Montesano, P. (1988). Facet joint injection in low-back pain. A prospective statistical study. *Spine* 13, 966–971.

Jung, J.-H., Kim, H.-I., Shin, D.-A., Shin, D.-G., Lee, J.-O., Kim, H.-J., Chung, J.-H. (2007). Usefulness of pain distribution pattern assessment in decision-making for the patients with lumbar zygapophyseal and sacroiliac joint arthropathy. *J Korean Med Sci* 22, 1048–1054.

Jüni, P., Witschi, A., Bloch, R., Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 282, 1054–1060.

Laslett, M., Öberg, B., Aprill, C., McDonald, B. (2004). Zygapophysial joint blocks in chronic low back pain: A test of Revel's model as a screening test. *BMC Musculoskelet Disord* 5, 43.

Laslett, M., McDonald, B., Tropp, H., Aprill, C., Öberg, B. (2005). Agreement between diagnoses reached by clinical examination and available reference standards: A prospective study of 216 patients with lumbopelvic pain. *BMC Musculoskelet Disord* 6, 28.

Laslett, M., McDonald, B., Aprill, C., Tropp, H., Őberg, B. (2006). Clinical predictors of screening lumbar zygapophyseal joint blocks: Development of clinical prediction rules. *Spine J* 6, 370–379.

Leeflang, M.M., Rutjes, A.W., Reitsma, J.B., Hooft, L., Bossuyt, P.M. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ* 185, E537–E544.

Lewinnek, G., Warfield, C. (1986). Facet joint degeneration as a cause of low back pain. *Clin Orthop Relat Res* 213, 216–222.

Manchikanti, L., Pampati, V., Fellows, B., Bakhit, C. (1999). Prevalence of lumbar facet joint pain in chronic low back pain. *Pain Physician* 2, 59–64.

Manchikanti, L., Pampati, V., Fellows, B., Baha, A. (2000a). The inability of the clinical picture to characterize pain from facet joints. *Pain Physician* 3, 158–166.

Manchikanti, L., Pampati, V., Fellows, B., Bakhit, C. (2000b). The diagnostic validity and therapeutic value of lumbar facet joint nerve blocks with or without adjuvant agents. *Curr Pain Headache Rep* 4, 337–344.

Manchikanti, L., Boswell, M., Singh, V., Pampati, V., Damron, K., Beyer, C. (2004). Prevalence of facet joint pain in chronic spinal pain of cervical, thoracic, and lumbar regions. *BMC Musculoskelet Disord* 5, 15.

Manchikanti, L., Singh, V., Datta, S., Cohen, S., Hirsch, J. (2009). Comprehensive review of epidemiology, scope, and impact of spinal pain. *Pain Physician* 12, 35–70.

Martin, B.I., Deyo, R.A., Mirza, S.K., Turner, J.A., Comstock, B.A., Hollingworth, W., Sullivan, S.D. (2008). Expenditures and health status among adults with back and neck problems. *JAMA* 299, 656–664.

Murtagh, F. (1988). Computed tomography and fluoroscopy guided anesthesia and steroid injection in facet syndrome. *Spine* 13, 686–689.

Raymond, J., Dumas, J., Lisbona, R. (1984). Nuclear imaging as a screening test for patients referred for intraarticular facet block. *J Can Assoc Radiol* 35, 291–292.

Reitsma, J., Glas, A., Rutjes, A., Scholten, R., Bossuyt, P., Zwinderman, A. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 58, 982–990.

Revel, M., Listrat, V., Chevalier, X., Dougados, M., N'guyen, M., Vallee, C., Wybier, M., Gires, F., Amor, B. (1992). Facet joint block for low back pain: Identifying predictors of a good response. *Arch Phys Med Rehab* 73, 824–828.

Revel, M., Poiraudeau, S., Auleley, G., Payan, C., Denke, A., Nguyen, M., Chevrot, A., Fermanian, J. (1998). Capacity of the clinical picture to characterize low back pain relieved by facet joint anesthesia: Proposed criteria to identify patients with painful facet joints. *Spine* 23, 1972–1976.

Riegelman, R.K. (2005). *Studying a Study and Testing a Test: How to Read the Medical Evidence* (Philadelphia, PA: Lippincott Williams & Wilkins).

Schwarzer, A., Aprill, C., Derby, R., Fortin, J., Kine, G., Bogduk, N. (1994a). Clinical features of patients with pain stemming from the lumbar zygapophysial joints. *Spine* 19, 1132–1137.

Schwarzer, A., Aprill, C., Derby, R., Fortin, J., Kine, G., Bogduk, N. (1994b). The false-positive rate of uncontrolled diagnostic blocks of the lumbar zygapophysial joints. *Pain* 58, 195–200.

Schwarzer, A., Derby, R., Aprill, C., Fortin, J., Kine, G., Bogduk, N. (1994c). Pain from the lumbar zygapophysial joints: A test of two models. *J Spinal Disord* 7, 331–336.

Schwarzer, A., Wang, S., O'Driscoll, D., Harrington, T., Bogduk, N., Laurent, R. (1995). The ability of computed tomography to identify a painful zygapophysial joint in patients with chronic low back pain. *Spine* 20, 907–912.

Smidt, N., Rutjes, A., Van der Windt, D., Ostelo, R., Bossuyt, P., Reitsma, J., Bouter, L., de Vet, H. (2006). The quality of diagnostic accuracy studies since the STARD statement Has it improved? *Neurology* 67, 792–797.

Sufka, A., Hauger, B., Trenary, M., Bishop, B., Hagen, A., Lozon, R., Martens, B. (1998). Centralization of low back pain and perceived functional outcome. *J Orthop Sports Phys* 27, 205–212.

Van der Windt, D., Simons, E., Riphagen, I., Ammendolia, C., Verhagen, A., Laslett, M., Devillé, W., Deyo, R., Bouter, L., de Vet, H. (2010). Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain. *Cochrane Database Syst Rev* 4, Art. No.: CD007431. DOI: 10.1002/14651858.CD007431.

Werneke, M., Hart, D. (2001). Centralization phenomenon as a prognostic factor for chronic low back pain and disability. *Spine* 26, 758–764.

Whiting, P., Rutjes, A., Reitsma, J., Bossuyt, P., Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 3, 25–37.

Whiting, P., Harbord, R., Kleijnen, J. (2005). No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol* 5, 19.

Whiting, P., Rutjes, A., Westwood, M., Mallett, S., Deeks, J., Reitsma, J., Leeflang, M., Sterne, J., Bossuyt, P. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 155, 529–536.

van Wijk, R., Geurts, J., Wynne, H., Hammink, E., Buskens, E., Lousberg, R., Knape, J., Groen, G. (2005). Radiofrequency denervation of lumbar facet joints in the treatment of chronic low back pain: A randomized, double-blind, sham lesion-controlled trial. *Clin J Pain* 21, 335–344.

Young, S., Aprill, C., Laslett, M. (2003). Correlation of clinical examination characteristics with three sources of chronic low back pain. *Spine J* 3, 460–465.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Search strategy.

Appendix S2. Criteria for Quality Assessment (QUADAS-2).

Appendix S3. Patient characteristics.

Appendix S4. Data extraction.