Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain (Protocol)

van der Windt DAWM, Simons E, Riphagen I, Ammendolia C, Verhagen AP, Laslett M, Devillé W, Aertgeerts B, Deyo RA, Bouter LM, de Vet HCW

THE COCHRANE
COLLABORATION®

This is a reprint of a Cochrane protocol, prepared and maintained by The Cochrane Collaboration and published in *The Cochrane Library* 2009, Issue 3

http://www.thecochranelibrary.com

WILEY
Publishers Since 1807

TABLE OF CONTENTS

[Diagnostic Test Accuracy Protocol]

Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain

Daniëlle AWM van der Windt[1], Emmanuel Simons[2], Ingrid Riphagen[3], Carlo Ammendolia[4], Arianne P Verhagen[5], Mark Laslett[6], Walter Devillé[7], Bert Aertgeerts[2], Rick A Deyo[8], Lex M Bouter[9], Henrica CW de Vet[10]

[1]Department of General Practice , EMGO Institute, Amsterdam, Netherlands. [2]Belgian Branch of the Dutch Cochrane Centre, CEBAM, Leuven, Belgium. [3]Medical Library , VU University Amsterdam, Amsterdam, Netherlands. [4]Rehabilitation Solutions, Toronto Western Hospital, University Health Network, Toronto, Canada. [5]Department of General Practice, Erasmus Medical Centre University , Rotterdam, Netherlands. [6]PhysioSouth, Christchurch, New Zealand. [7]Ingternational and Migrant Health, NIVEL, Utrecht, Netherlands. [8]Evidence-Based Family Medicine, Oregon Health and Science University, Portland, OR, USA. [9]Executive Board of VU University Amsterdam, Amsterdam, Netherlands. [10]Department of Epidemiology and Biostatistics, EMGO Institute, Amsterdam, Netherlands

Contact address: Daniëlle AWM van der Windt, Department of General Practice , EMGO Institute, VU University Medical Center , Van der Boechorststraat 7, Amsterdam, 1081 BT, Netherlands. dawm.vanderwindt@vumc.nl. d.van.der.windt@cphc.keele.ac.uk. (Editorial group: Cochrane Back Group.)

Cochrane Database of Systematic Reviews, Issue 3, 2009 (Status in this issue: *Unchanged*)
Copyright © 2009 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.
DOI: 10.1002/14651858.CD007431
This version first published online: 8 October 2008 in Issue 4, 2008. (Help document - Dates and Statuses explained)

This record should be cited as: van der Windt DAWM, Simons E, Riphagen I, Ammendolia C, Verhagen AP, Laslett M, Devillé W, Aertgeerts B, Deyo RA, Bouter LM, de Vet HCW. Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain. *Cochrane Database of Systematic Reviews* 2008, Issue 4. Art. No.: CD007431. DOI: 10.1002/14651858.CD007431.

ABSTRACT

This is the protocol for a review and there is no abstract. The objectives are as follows:

The general aim of our review is to provide information that may assist the clinician in making decisions about appropriate management in patients with low-back pain and leg pain suspected of having radicular pain and radiculopathy due to disc herniation. More specifically, the objective of this systematic review is to assess the diagnostic performance of tests performed during physical examination in the identification of radicular pain and radiculopathy due to lumbar disc herniation in patients with low-back and leg pain.

The secondary objective of this review is to assess the influence of sources of heterogeneity on the diagnostic accuracy of tests performed during physical examination, in particular the type of reference standard, health care setting, spectrum of disease, and study design.

BACKGROUND

Low-back pain (LBP) is a common cause of disability in Western industrialised countries. Although many people experience at least one episode of low-back pain in their life, in up to 85% of the patients, no specific pathology is identified (Deyo 1992). In patients who report symptoms radiating into the lower leg (sciatica), clinicians evaluate the likelihood of radiculopathy due to lumbar disc herniation through history and physical examination. The most commonly used physical tests include the straight leg raising test, crossed straight leg raising test, tendon reflexes, and signs of paresis, atrophy or sensory deficits (Deyo 1992; Rebain 2002; Rebain 2003; van den Hoogen 1995). Of all patients who experience episodes of back pain, fewer than 2% will undergo surgery for suspected disc herniation (Deyo 1990).

In patients with LBP, physicians or therapists use the information gained during history and physical examination to decide on a management plan. Part of this management plan includes making decisions about diagnostic imaging, or the potential value of surgical intervention. Certain findings of physical examination (for example, positive straight leg raising test) may predict better outcomes of surgery and chemonucleolysis (Kim 2002; Kohlboeck 2004).

Several systematic reviews have summarised the results of available studies on the diagnostic performance of the physical examination for the identification of lumbar radiculopathy in these patients (Deyo 1992; Deville 2000a; van den Hoogen 1995; Vroomen 1999). Three of these reviews included an assessment of the methodological quality of primary diagnostic studies (Deville 2000a; Deyo 1992; Vroomen 1999), and two offered a quantitative summary of the findings (Deville 2000a; Vroomen 1999). These systematic reviews show that most physical tests may have adequate sensitivity, but poor specificity in the identification of disc herniation, while some tests have high specificity and low sensitivity. The diagnostic accuracy varied considerably across primary diagnostic studies. Several factors may explain this heterogeneity, including variation in patient populations (health care setting, spectrum of disease), variation in the way physical tests were performed, differences in the selection of the reference standard, or differences in study design. Potential sources of heterogeneity were only explored by Devillé et al (Deville 2000a). The results suggested that the diagnostic accuracy of the straight leg raising test decreased with a more valid study design, a more homogeneous case-mix, and with more recent publication. However, this systematic review examined only the straight leg raising test, and needs updating. This systematic review aims to provide updated evidence on the diagnostic performance of several tests carried out during physical examination, includes an assessment of methodological quality, and pays specific attention to the potential influence of sources of heterogeneity.

OBJECTIVES

The general aim of our review is to provide information that may assist the clinician in making decisions about appropriate management in patients with low-back pain and leg pain suspected of having radicular pain and radiculopathy due to disc herniation. More specifically, the objective of this systematic review is to assess the diagnostic performance of tests performed during physical examination in the identification of radicular pain and radiculopathy due to lumbar disc herniation in patients with low-back and leg pain.

Investigation of sources of heterogeneity

The secondary objective of this review is to assess the influence of sources of heterogeneity on the diagnostic accuracy of tests performed during physical examination, in particular the type of reference standard, health care setting, spectrum of disease, and study design.

METHODS

Criteria for considering studies for this review

Types of studies

We will consider primary diagnostic studies if they compare the results of tests performed during physical examination in the identification of radicular pain and radiculopathy due to lumbar disc herniation with those of a reference standard. Only cohort studies and case-control studies will be included in the review. We will only include results from full reports. If studies have been reported in abstracts or conference proceedings we will search for full publications.

Participants

We will include studies that assess diagnostic accuracy of physical examination in back pain patients with symptoms suspected to be caused by lumbar disc herniation. We will include studies carried out in primary as well as secondary care, and examine the potential influence of the setting on diagnostic performance. We will exclude studies that enrol patients who have been previously diagnosed with other specific causes of low-back pain (for example, infection, tumour, severe osteoarthritis, or fractures).

Index tests

Studies on all relevant physical examination tests will be eligible for inclusion, such as the straight leg raising test (test of Lasègue), crossed straight leg raising test, femoral nerve stretch test, depressed reflexes, atrophy, paresis or sensory deficits. We will include studies in which the diagnostic performances of individual aspects of the physical examination are evaluated separately, or in combination.

In the case of a combination, the study should clearly describe which tests are included in the combination, and how. We will not include studies in which a "clinical diagnosis" (some unknown combination of history and physical examination) is compared with the results of a reference standard.

Target conditions

We will select diagnostic studies if the aim of the diagnostic test is to identify radicular pain and radiculopathy due to lumbar disc herniation. We will exclude studies in which other specific causes of low-back pain (for example, infection, tumour, severe osteoarthritis, or fractures) are the target condition, and diagnostic testing is aimed at identifying these conditions.

Reference standards

We will include studies if physical examination is compared to 1) diagnostic imaging: Magnetic Resonance Imaging (MRI), Computed Tomography (CT), myelography; or 2) findings at surgery. The quality of diagnostic imaging as a reference test has been debated as herniated discs can be found on diagnostic imaging in 20% to 30% of symptom-free persons (Boden 1990). Therefore, separate (stratified) analyses will be carried out for these two different reference standards. Surgical findings as a reference standard have the disadvantage that only patients with a strong suspicion of radiculopathy will be subjected to surgery (risk of verification bias).

Search methods for identification of studies

Electronic searches

A search strategy has been developed in collaboration with a medical information specialist (IR). We will search relevant computerised databases for eligible diagnostic studies: MEDLINE, OLDMEDLINE, EMBASE, and CINAHL (all publications until present). The search strategy for MEDLINE is presented in Appendix 1 and has been adapted for EMBASE and CINAHL. A previous systematic review on the diagnostic performances of the straight leg raising test has been used as a point of reference (Deville 2000a). All publications included in that review are indexed in MEDLINE. The search was refined until all publications in the review were identified by our search. The strategy uses several combinations of searches related to the patient population, aspects of physical examination, and the target condition. A methodological filter for the identification of primary diagnostic studies (search 4c) has been added to one of the searches to increase the specificity of the search, and to limit the harvest to less than 2000 hits. This filter is highly sensitive and partly based on those proposed by Devillé et al (Deville 2000b), and Bachman et al (Bachmann 2002; Bachmann 2003).

Searching other resources

We will check the reference lists of all retrieved relevant publications (primary diagnostic studies). An additional electronic search was composed to identify relevant (systematic) reviews in MEDLINE and Medion (www.mediondatabase.nl), and their references will be checked. In addition, we will contact researchers in the field of low-back pain research to identify additional diagnostic studies. No language restrictions will be applied.

Data collection and analysis

Selection of studies

Two review authors (BA and ES) will independently apply the selection criteria to all citations (titles and abstracts) identified by the search strategy described above. Consensus meetings will be organised to discuss any disagreement regarding selection. Final selection will be based on a review of full publications, which will be retrieved for all studies that either meet the selection criteria, or for which there will be uncertainty regarding selection. A third review author (DvdW) will be consulted in cases of persisting disagreement.

Data extraction and management

For each included study, we will use a standardised form to extract characteristics of participants, the index tests and reference standard, and aspects of study methods.

- Characteristics of participants will include setting (primary / secondary care); inclusion and exclusion criteria; enrolment (consecutive or non-consecutive); number of subjects (including number eligible for the study, number enrolled in the study, number receiving index test and reference standard, number for whom results are reported in the two-by-two table, reasons for withdrawal); duration and history of low-back pain, and presence of leg pain.
- Test characteristics will include the type of test, methods of execution, experience and expertise of the assessors, type of reference standard, and cut-off points for diagnosing radicular pain and radiculopathy due to lumbar disc herniation. Positivity thresholds (interpretations of "positive" results) may vary across studies, and some studies may present diagnostic performance of an index test at several different cut-off points. We will extract data regarding the most commonly used cut-off points.
- Aspects of study methods will include the basic design of the study (case-control, prospective cohort, or historical cohort with data collection based on medical records), time and treatment between index test and reference standard, and quality criteria (Table 1).

Table 1. Items for the quality assessment of diagnostic accuracy studies (QUADAS)

Item and Guide to classification
1. Was the spectrum of patients representative of the patients who will receive the test in practice? Is it a selective sample of patients? Differences in demographic or clinical features between the study population and the source population may lead to selection bias or spectrum variation. In this item we will focus on selection bias: is a selective sample of patients included? • **Classify as 'yes'** if a consecutive series of patients or a random sample has been selected. Information should be given about setting, in- and exclusion criteria, and preferably number of patients eligible and excluded. If a mixed population of primary and secondary care patients is used: the number of participants from each setting is presented. • **Classify as 'no'** if healthy controls are used. Score also 'no' if non-response is high and selective, or there is clear evidence of selective sampling. Score also 'no' if a population is selected that is otherwise unsuitable, for example, patients are known to have other specific causes of LBP (severe OA, malignancies, etc). • **Classify as 'unclear**' if insufficient information is given on the setting, selection criteria, or selection procedure to make a judgment.
2. Is the reference standard likely to classify the target condition correctly? Estimates of test performance are based on the assumption that the reference standard will identify nerve root compression due to disc herniation with 100% sensitivity and 100% specificity. Such reference standards are rare. Errors due to an imperfect reference standard may bias the estimation of diagnostic performance. For this review acceptable reference standards are: 1) findings at surgery demonstrating nerve root compression or irritation due to disc herniation; and 2) myelography indicating nerve root compression; and 3) although probably of lower quality, CT/MRI findings indicating nerve root compression; • **Classify as 'yes'** if one of these procedures is used as reference standards. • **Classify as 'no'** if you seriously question the methods used, if consensus among observers, or a combination of aspects of physical examination and history ('clinical judgement') is used as reference standard. *(Use of imaging/surgery is actually a selection criterion, so the latter may not occur)* • **Classify as 'unclear'** if insufficient information is given on the reference standard. • **Classify as 'not able'** if you consider yourself not capable to assess this item. If you have doubts, for example, regarding the quality of MRI-procedures but feel not competent to make an adequate assessment, we can consult a radiologist.
3. Is the time period between the reference standard and the index test short enough to be reasonably sure that the target condition did not change between the two tests? The index tests and reference standard should ideally be carried out at the same time. If there is a considerable delay, misclassification (due to spontaneous recovery or worsening of the condition) may occur. • **Classify as 'yes'** if the time period between physical examination and the reference standard is one week or less. • **Classify as 'no'** if the time period between physical examination and the reference standard is longer than one week. • **Classify as 'unclear**' if there is insufficient information on the time period between index tests and reference standard.
4. Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis? When not all of the study patients receive confirmation of their diagnosis by a reference standard, partial verification bias may occur. Bias is very likely if the results of the index test influence the decision to perform the reference standard. Random allocation of patients to the reference standard should in theory not affect diagnostic performance. [Verification bias is also known as work-up bias or sequential ordering bias]. • **Classify as 'yes'** if it is clear that all patients who received the index test went on to receive a reference standard, even if the reference standard is not the same for all patients. • **Classify as 'no'** if not all patients who received the index test received verification by a reference standard. • **Classify as 'unclear'** if insufficient information is provided to assess this item.

Table 1. Items for the quality assessment of diagnostic accuracy studies (QUADAS) (*Continued*)

5. Did patients receive the same reference standard regardless of the index test result?

Differential verification bias occurs when the results of the index tests are verified by different reference standards. This is not unlikely in this review: some patients may be referred for surgery following physical examination, whereas others only go on to receive diagnostic imaging. Bias is likely to occur when this decision depends on the results of the index test.

- **Classify as 'yes'** if it is clear that all patients receiving the index test are subjected to the same reference standard.
- **Classify as 'no'** if different reference standards are used.
- **Classify as 'unclear'** if insufficient information is provided to assess this item.

6. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?

It is not unlikely that the results of the physical examination are used when establishing the final diagnosis. In this case incorporation bias may occur (overestimating diagnostic accuracy). Knowledge of the results of the index test does not necessarily mean that these results are incorporated in the reference standard. For example, if the reference standard consists of MRI-results only (regardless of knowledge of the results of the straight leg raising test), the index test is *not* part of the reference standard. However, if the final diagnosis is based on the results of both MRI-findings *and* a positive straight leg raising test, incorporation bias will occur.

- **Score 'yes'** if the index is no part of the reference standard.
- **Score 'no'** if the index test is clearly part of the reference standard.
- **Score 'unclear'** if insufficient information is provided to assess this item.

7. Were the index test results interpreted without knowledge of the results of the reference standard?

8. Were the reference standard results interpreted without knowledge of the results of the index test?

Interpretation of the results of physical examination may be influenced by knowledge of the results of the reference standard, and vice versa. This is known as reviewer bias, and may lead to over-estimation of diagnostic accuracy. In our review the risk of bias may be substantial as both index test and reference standard often involve a subjective assessment of results. If the index test always precedes the reference standard, interpretation of the results of the index test will usually be without knowledge of the results of the reference standard. The reverse may also be true, although surgery is unlikely to precede physical examination.

- **Classify as 'yes'** if the test results (index or reference standard) are interpreted blind to the results of the other test. Score also **'yes'** if the sequence of testing is always the same and, consequently, one of the test is interpreted blind for the other.
- **Classify as 'no'** if the assessor is aware of the results of the other test.
- **Classify as 'unclear'** if insufficient information is given on independent or blind assessment of the index test or reference standard.

9. Were the same clinical data available when the index test results were interpreted as would be available when the test is used in practice?

The knowledge of demographic and clinical data, such as age, gender, symptoms, history of low-backpain, previous treatments, or other aspects of physical examination may influence the interpretation of test results. The way this item is scored depends on the objective of the index test. If an aspect of physical examination is intended to replace other tests, these clinical data should *not* be available. However, if in practice clinical data are usually available when interpreting the results of the index test, this information should be available to the assessors of the index test.

- **Classify as 'yes'** if clinical data (i.e. patient history, other physical tests) would normally be available when the test results are interpreted and similar data are available in the study.
- **Classify as 'yes'** if clinical data would normally not be available when the test results are interpreted and these data are also not available in the study.
- **Classify as 'no'** if this is not the case, e.g. if other test results are available that can not be regarded as part of routine care.
- **Classify as 'unclear'** if the paper does not explain which clinical information was available at the time of assessment.

Table 1. Items for the quality assessment of diagnostic accuracy studies (QUADAS) (*Continued*)

10. Were uninterpretable / intermediate test results reported?

Uninterpretable or intermediate test results are often not reported in diagnostic studies. Authors may simply remove these results from the analysis, which may lead to biased results of diagnostic performance. If uninterpretable or intermediate test results occur randomly and are not related to disease status, bias is unlikely. Whatever the cause of uninterpretable results they should be reported in order to estimate their potential influence on diagnostic performance.

- **Classify as 'yes'** if all test results are reported for all patients, including uninterpretable, indeterminate or intermediate results.
- **Classify as 'yes'** if the authors do not report any uninterpretable, indeterminate or intermediate results AND the results are reported for all patients who were described as having been entered into the study.
- **Classify as 'no'** if you think that such results occurred, but have not been reported.
- **Classify as 'unclear'** if it is unclear whether all results have been reported.

11. Were withdrawals from the study explained?

Patients may withdraw from the study before the results of both index test and reference standard are known. If withdrawals systematically differ from patients remaining in the study, then estimates of diagnostic test performance may be biased. A flow chart is sometimes provided (in more recently published papers) which may help to score this item.

- **Classify as 'yes'** if it is clear what happens to all patients who entered the study (all patients are accounted for, preferably in a flow chart).
- **Classify as 'yes'** if the authors do not report any withdrawals AND if the results are available for all patients who were reported to have been entered in the study.
- **Classify as 'no'** if it is clear that not all patients who were entered completed the study (received both index test and reference standard), and not all patients are accounted for.
- **Classify as 'unclear'** when the paper does not clearly describe whether or not all patients completed all tests, and are included in the analysis.

Note: In many diagnostic studies one may doubt whether or not all eligible patients have been entered in the study and are described in the paper. This issue is more strongly related to selection bias and will be scored under item 1.

Additional QUADAS items

12. Did the study provide a clear definition of what was considered to be a "positive" result of the index test?

Aspects of physical examination, for example the straight leg raising test, require a subjective judgement. Furthermore, several methods of performing the test have been described, and several cut-offs have been proposed. Consequently, it is essential that an adequate description is given of the methods used to carry out (aspects of) physical examination, and how a positive result is defined.

- **Classify as 'yes'** if the paper provides a clear description of the way the index test is performed, including a definition of a positive test result.
- **Classify as 'no'** if no description is given of the way the index test is performed, and no definition is given of a positive test result.
- **Classify as 'unclear'** if the methods of the index test are described, but no clear definition of a positive result has been provided, or vice versa.

13. Was treatment withheld until both index test and reference standard were performed?

If index tests and reference standard are not performed on the same day, some type of intervention may be initiated in between index test and reference standard. This might lead to misclassification (if some recovery of symptoms occurs).

- **Classify as 'yes'** if no treatment is given in the time period between physical examination and the reference standard.

Table 1. Items for the quality assessment of diagnostic accuracy studies (QUADAS) *(Continued)*

• **Classify as 'no'** if an intervention is given that in your opinion could possibly influence the prognosis of low-backpain due to nerve root compression / irritation. • **Classify as 'unclear'** if there is insufficient information regarding treatment between index test and reference standard.
14. Were data on observer variation reported and within acceptable range? Studies on the reproducibility of physical examination in patients with musculoskeletal pain show that there may be considerable inter-observer variation. This may strongly influence the diagnostic performance of the index test. It is difficult to give minimal cut-off scores for inter-observer agreement. A kappa or ICC of 0.70 is often considered to be acceptable, but this is certainly an arbitrary definition. • **Classify as 'yes'** if the paper provides information on inter-observer variation, and the results are acceptable. • **Classify as 'no'** if information is given on inter-observer variation, and the results demonstrate poor agreement. • **Classify as 'unclear'** if there is insufficient information is provided regarding inter-observer variation.

Each item is classified as "yes" (adequately addressed); "no" (inadequately addressed) or "unclear" (inadequate detail presented to allow a judgement to be made).

We will extract the diagnostic two-by-two table (true positive, false positive, true negative, and false negative rates) from the publications, or if not available, reconstruct the two-by-two table using information on relevant parameters (sensitivity, specificity or predictive values). Eligible studies for which the diagnostic two-by-two table cannot be reconstructed will be presented in the review, but will not be included in the quantitative analyses. Two review authors will independently extract the data (ML and DvdW) to ensure adequate reliability of collected data. For each study, we will present aspects of study design, characteristics of the population, index test, reference standard and diagnostic parameters (sensitivity and specificity) in tables. We will use two primary diagnostic studies not included in the review (on the diagnostic accuracy of physical examination in patients with shoulder pain) to pilot the data extraction form.

Assessment of methodological quality

The quality of each study will be assessed by at least two review authors (AV, CA, DvdW), using the Quality Assessment of Diagnostic Accuracy Studies list (QUADAS; Whiting 2004). The Cochrane Methods Group on Meta-analysis of Diagnostic and Screening Tests recommends these items (Table 1) (Handbook 2005). The QUADAS tool consists of 11 items that refer to internal validity (for example, blind assessment of index and reference test, or avoidance of verification bias). Three additional items described in the Cochrane Handbook for Diagnostic Test Accuracy Reviews (Handbook 2005) are of relevance to this review and will also be scored. These additional items refer to the definition of the positivity threshold of the index test, treatment given between index test and reference standard, and observer variation (Table 1).

The review author s will classify each item as "yes" (adequately addressed); "no" (inadequately addressed) or "unclear" (inadequate detail presented to allow a judgement to be made). Guidelines for the assessment of each item will be made available to the review authors (see Table 1). Again, assessment of methodological quality will be pre-tested using two studies not included in the review. We will quantify inter-observer agreement by computing the percentage agreement for each item of the checklist. Disagreements will be resolved by consensus and, if necessary, by third party (HdV) adjudication.

We will not apply weights to the different items and will not use a summary quality score to incorporate studies with certain levels of quality in the analysis. We will explore the influence of negative classification of important items using subgroup analyses or meta-regression analyses (see below). The following items will be considered for these analyses: item 1 (spectrum variation / selective sample), item 4 (verification bias), items 7 and 8 (blinded interpretation of index test and reference standard) and item 11 (explanation of withdrawals).

Statistical analysis and data synthesis

The two key and commonly reported parameters of diagnostic test accuracy are sensitivity and specificity. As there is a trade-off between these two parameters, we will analyse them jointly. Sensitivities and specificities for each index test with 95% confidence intervals will be presented in tables. In addition, a scatterplot of study-specific estimates of sensitivity and 1-specificity will be used to display the data in Receiver Operating Characteristic (ROC) space.

Summary ROC analysis

Summary ROC analysis characterises the relationship between sensitivity and 1-specificity across studies and takes into account variation in the threshold for test positivity between studies. Littenberg and Moses (Littenberg 1993) proposed a method of generating a summary ROC curve using simple linear regression that is frequently used (Deeks 2001; Deville 2002; Glas 2003a; Handbook 2005; Littenberg 1993). In this method, pairs of sensitivity and specificity are transformed to log odds (logit) scales. Next, a linear regression equation is estimated using the transformed data: $D = \alpha + \beta S$, with D = logit (sensitivity) - logit (1-specificity), and S = logit (sensitivity + logit(1-specificity)). S can be considered to be a proxy of the positivity cut-off point, and S will increase when the overall proportion of positive test results increases. D represents the natural logarithm of the Diagnostic Odds Ratio (DOR) (Littenberg 1993). The DOR combines both sensitivity and specificity into one parameter of diagnostic accuracy, presenting the ratio of the odds of a positive test result in the diseased group to the odds of a positive test in the non-diseased group (Glas 2003b). The model shows how test accuracy (measured by the DOR) varies with a proxy of the positivity threshold across studies. When $\beta = 0$, the DOR for each study does not depend on the cut-off point parameter, and α is then the log of the diagnostic odds ratio. When β 0, the DOR varies with the threshold (Handbook 2005; Littenberg 1993).

The Moses and Littenberg model departs from a fixed-effect model, and does not consider the possibility of random variation across studies. Two statistically rigorous methods for the meta-analysis of data from diagnostic test accuracy studies: hierarchical SROC analysis (Rutter 2001) and bivariate analyses (Reitsma 2005), enable the use of a random-effects model. A random effects model allows the review authors to take into account variability both within and between studies. The HSROC model assumes that there is an underlying ROC curve in each study with parameters α and β that characterise the accuracy and asymmetry of the curve, as in the Moses and Littenberg model, but the parameters α and the positivity threshold are assumed to vary between studies: both are assumed to have normal distributions as in conventional random-effects meta-analysis. Separate equations are defined for within-study variation ('Level I') and between-study variation ('Level II') (Handbook 2005; Rutter 2001; Harbord 2007). We will use this method to compute and plot hierarchical SROC curves for subsets of studies showing sufficient clinical homogeneity (same reference standard, similar definition of disc herniation).

Bivariate analysis

Rather than using a single outcome measure per study (the DOR), the bivariate model preserves the two-dimensional nature of diagnostic data by directly analysing the logit transformed sensitivity and specificity of each study in a single model. This model incorporates the correlation that might exist between sensitivity and specificity within studies caused by possible differences in threshold between studies (Harbord 2007).

The model produces the following results: a random-effects estimate of the mean sensitivity and specificity with corresponding 95% confidence intervals, the amount of between-study variation for sensitivity and specificity separately, and the strength and shape of the correlation between sensitivity and specificity. Using these results, we will calculate a 95% confidence ellipse (i.e. bivariate confidence interval) around the summary estimate of sensitivity and specificity. All the results will be transformed back to the original scale, and plotted in ROC space (Harbord 2007). We will only use bivariate analysis to present pooled estimates of sensitivity and specificity if studies show sufficient clinical homogeneity (same reference standard, similar definition of disc herniation), and results of sensitivity and specificity show sufficient statistical homogeneity. All meta-analyses will be carried out using STATA software.

Investigations of heterogeneity

Several factors (next to variability in the positivity threshold) may contribute to heterogeneity in diagnostic performance across studies. We will investigate the potential influence of differences in the type of reference standard (surgery versus imaging); study population (primary versus secondary care, previous lumbar disc surgery), and study design (prospective cohort or other designs, scores on items 1, 4, 7, 8, and 11 of the QUADAS checklist). The HSROC or bivariate models can be extended to include study level covariates (characteristics of individual studies) (Handbook 2005). This will allow a statistical assessment of the evidence for an association between test accuracy and potential sources of heterogeneity (meta-regression analysis). Approaches to study the potential influences of more than one study level covariate are only feasible if a large number of studies report on the same index test, and provide sufficient information on the covariates of interest. If this is not the case, we will use subgroup analyses to study the potential influence of sources of heterogeneity.

Finally, we will summarise the findings of the review in a summary table (Handbook 2005), which includes a summary estimate of sensitivity, specificity, and likelihood ratios for relevant tests and subgroups of studies (for example, studies on patients in primary or secondary care, and studies using different reference standards). The presentation of this summary table will make diagnostic information more accessible to healthcare providers and other end users.

ACKNOWLEDGEMENTS

We would like to thank Petra Macaskill, Rob Scholten and Vicki Pennick for their useful comments on the first version of the protocol.

REFERENCES

Additional references

Bachmann 2002

Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *Journal of the American Medical Informatics Association* 2002;**9**(6):653–8.

Bachmann 2003

Bachmann LM, Estermann P, Kronenberg C, ter Riet G. Identifying diagnostic accuracy studies in EMBASE. *Journal of the Medical Library Association* 2003;**91**(3):341–6.

Boden 1990

Boden SD, Davis DO, Dina TS, Patronas NJ, Wiesel SW. Abnormal magnetic-resonance scans of the lumbar spine in asymptomatic subjects. A prospective investigation. *Journal of Bone and Joint Surgery* 1990;**72**(3):403–8.

Deeks 2001

Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Davey Smith G, Altman DG editor(s). *Systematic reviews in health care*. London: BMJ Publishing Group, 2001:248–82.

Deville 2000a

Devillé WL, van der Windt DA, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000;**25**(9):1140–7.

Deville 2000b

Devillé WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *Journal of Clinical Epidemiology* 2000;**53**(1):65–9.

Deville 2002

Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al.Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Medical Research Methodology* 2002;**2**:9.

Deyo 1990

Deyo RA, Loeser JD, Bigos SJ. Herniated lumbar intervertebral disk. *Annals of Internal Medicine* 1990;**112**(8):598–603.

Deyo 1992

Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain?. *JAMA* 1992;**268**(6):760–5.

Glas 2003a

Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. *Journal of Urology* 2003;**169**(6):1975–82.

Glas 2003b

Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 2003;**56**(11):1129–35.

Handbook 2005

Cochrane Handbook for Diagnostic Test Accuracy Reviews. July 2005.

Harbord 2007

Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studiesy. *Biostatistics* 2007;**8**(2):239–51.

Kim 2002

Kim YS, Chin DK, Yoon DH, Jin BH, Cho YE. Predictors of successful outcome for lumbar chemonucleolysis: analysis of 3000 cases during the past 14 years. *Neurosurgery* 2002;**51**(5 Suppl):S123–8.

Kohlboeck 2004

Kohlboeck G, Greimel KV, Piotrowski WP, Leibetseder M, Krombholz-Reindl M, Neuhofer R, et al.Prognosis of multifactorial outcome in lumbar discectomy: a prospective longitudinal study investigating patients with disc prolapse. *Clinical Journal of Pain* 2004;**20**(6):455–61.

Littenberg 1993

Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making* 1993;**13**(4):313–21.

Rebain 2002

Rebain R, Baxter GD, McDonough S. A systematic review of the passive straight leg raising test as a diagnostic aid for low back pain (1989 to 2000). *Spine* 2002;**27**(17):E388–95.

Rebain 2003

Rebain R, Baxter GD, McDonough S. The passive straight leg raising test in the diagnosis and treatment of lumbar disc herniation: a survey of United kingdom osteopathic opinion and clinical practice. *Spine* 2003;**28**(15):1717–24.

Reitsma 2005

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005;**58**(10):982–90.

Rutter 2001

Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001;**20**(19):2865–84.

van den Hoogen 1995

van den Hoogen HM, Koes BW, van Eijk JT, Bouter LM. On the accuracy of history, physical examination, and erythrocyte sedimentation rate in diagnosing low back pain in general practice. A criteria-based review of the literature. *Spine* 1995;**20**(3):318–27.

Vroomen 1999

Vroomen PC, de Krom MC, Knottnerus JA. Diagnostic value of history and physical examination in patients suspected of sciatica due to disc herniation: a systematic review. *Journal of Neurology* 1999;**246**(10):899–906.

Whiting 2004

Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technology Assessment* 2004;**8**(25):iii, 1-234.

References to other published versions of this review

van der Windt 2008

van der Windt D, Simons E, Riphagen I, Ammendolia C, Verhagen A, Laslett M, Devillé W, Aertgeerts B, Deyo R, Bouter L, de Vet H. Physical examination for the diagnosis of lumbar radiculopathy due to disc herniation in patients with low-back pain. *Cochrane Database of Systematic Reviews* 2007, Issue 4. [: Chichester, UK: Wiley. Available from: http://www3.interscience.wiley.com/cgi–bin/mrwhome/106568753/DTAP7.pdf]

* *Indicates the major publication for the study*

APPENDICES

Appendix 1. MEDLINE search strategy

1 Index test: tests performed during physical examination

1a

“straight leg raising”[tw] OR lasegue[tw] OR (provocation[tw] AND “intra abdominal pressure”[tw]) OR “bell test”[tw] OR “hyper-extension test”[tw] OR “femoral nerve stretch test”[tw] OR (achilles[tw] AND (areflexia[tw] OR reflex*[tw])) OR (knee[tw] AND (extens*[tw] OR reflex[tw])) OR “Reflex, stretch”[mesh] OR (dermatom*[tw] AND (somatosensory[tw] OR sensibility[tw])) OR slump[tw] OR (“muscle strength”[tw] AND leg[tw] AND (test[tw] OR tests[tw] OR testing[tw] OR sign[tw])) OR ((Bragard*[tw] OR Naffziger*[tw]) AND (test[tw] OR tests[tw] OR testing[tw] OR sign[tw])) OR (measur*[tw] AND “calf wasting”[tw]) OR (impair*[tw] AND “ankle reflex”[tw]) OR (weakness[tw] AND dorsiflexion[tw] AND foot[tw])

1b

Physical examination[mesh] OR “physical examination” OR “function test” OR “physical test” OR (clinical[tw] AND (diagnosis[tw] OR sign[tw] OR signs[tw] OR significance[tw] OR symptom*[tw] OR parameter*[tw] OR assessment[tw] OR finding*))

2. Population: low-back**pain and anatomical location**

2a

back pain[mesh] OR sciatica[mesh] OR “back ache”[tw] OR backache[tw] OR “back pain”[tw] OR dorsalgia[tw] OR lumbago[tw] OR sciatica[tw] OR ischias[tw] OR ischialgia[tw] OR lumboischialgia[tw] OR radiculalgia[tw] OR ((Pain[mesh] OR pain[tw] OR ache*[tw] OR aching[tw] OR complaint*[tw] OR dysfunction*[tw] OR disabil*[tw] OR neuralgia[tw]) AND (Back[mesh] OR spine[mesh] OR back[ti] OR lowback[tw] OR lumbar[tw] OR lumbal[tw] OR lumbo*[tw] OR sciatic[tw] OR spine[tw] OR spinal[tw] OR radicular[tw] OR “nerve root”[tw] OR “nerve roots”[tw] OR disk[tw] OR disc[tw] OR disks[tw] OR discs[tw] OR vertebra*[tw] OR intervertebra*[tw] OR sacroilia*[tw] OR Sacroiliac-joint[mesh]))

2b

low[tw] OR lower[tw] OR lowback[tw] OR sciatic*[tw] OR ischia*[tw] OR lumbo*[tw] OR lumba*[tw] OR sacroilia*[tw]

3. Target condition: lumbar radiculopathy

Intervertebral disk displacement[mesh] OR polyradiculopathy[mesh] OR radiculopath* OR radiculiti* OR ((disc OR discs OR disk OR disks) AND (displacement OR hernia* OR protru* OR avulsion*)) OR ((“nerve root” OR “nerve roots”) AND (compress* OR entrap* OR inflammat* OR disorder*)) OR (nerve compression syndromes[mesh] AND (Back[mesh] OR spine[mesh] OR back[ti] OR lowback[tw] OR lumbar[tw] OR lumbal[tw] OR lumbo*[tw] OR sciatic[tw] OR spine[tw] OR spinal[tw] OR radicular[tw] OR “nerve root”[tw] OR “nerve roots”[tw] OR disk[tw] OR disc[tw] OR disks[tw] OR discs[tw] OR vertebra*[tw] OR intervertebra*[tw] OR sacroilia*[tw] OR Sacroiliac-joint[mesh]))

4. Methodological filter (primary diagnostic studies)

4a

diagnosis[sh] OR pathophysiology[sh] OR etiology[sh]

4b

diagnosis[sh] OR diagnosis[mesh:noexp]

4c

Diagnostic errors[mesh] OR “Diagnosis, differential”[mesh] OR “Reproducibility of results”[mesh] OR Reference standards[mesh] OR “Sensitivity and specificity”[mesh] OR Comparative study[mesh] OR Evaluation studies[mesh] OR Evaluation studies[pt] OR

Longitudinal studies[mesh] OR sensitivit* OR specificit* OR accura* OR likelihood ratio* OR predict* OR index test OR reference test OR (false[tw] AND (positive[tw] OR negative[tw])) OR pretest[tw] OR pre-test[tw] OR posttest[tw] OR post-test[tw] OR “gold standard” OR roc[tw] OR odds[tw] OR validity OR validation OR validate* OR validation studies[pt] OR verif*[ti] OR evaluat*[ti] OR value*[ti] OR reference values[mesh] OR cutoff OR cut-off OR repeatability OR reproducibility OR efficacy OR reliability OR error*[tw] OR suitability[tw] OR utility[tw]

5. Exclusion criteria: children, reviews, case reports, animal studies

((child[mesh] OR infant[mesh]) NOT (adult[mesh] OR adolescent[mesh])) OR Review[pt] OR case reports[pt] OR (animals[mesh] NOT humans[mesh])

Searches (combinations)

A. 1a and (2a or 3) and 2b not 5
B. 1a and ((2a and 4a) or (3 and 4b)) not 5
C. 1b and 2a and 2b and 3 and (4a or 4b) not 5
D. 1b and 2b and 3 and 4b and 4c not 5
Final search: A or B or C or D

HISTORY

Protocol first published: Issue 4, 2008

CONTRIBUTIONS OF AUTHORS

DvdW checked selection of papers, contributed to quality assessment, carried out statistical analysis, and wrote the protocol and the review. IR designed the search strategy. ES and BA selected abstracts and papers. AV and CA carried out quality assessment, ML and DvdW carried out data extraction. LB and HdV provided methodological advice, RD provided methodological and clinical advice. All co-authors commented on several drafts of the protocol and review.

DECLARATIONS OF INTEREST

There are no conflicts of interest.