REVIEW ARTICLE

# Computed tomography for the diagnosis of lumbar spinal pathology in adult patients with low back pain or sciatica: a diagnostic systematic review

Rogier M. van Rijn · Merel Wassenaar · Arianne P. Verhagen · Raymond W. J. G. Ostelo ·
Abida Z. Ginai · Michiel R. de Boer · Maurits W. van Tulder · Bart W. Koes

**Abstract**

*Aim*  In low back pain if serious pathology is suspected diagnostic imaging could be performed. One of the imaging techniques available for this purpose is computed tomography (CT), however, insight in the diagnostic performance of CT is unclear.

*Method*  Diagnostic systematic review. Studies assessing the diagnostic accuracy of CT in adult patients suggested having low back pain caused by specific pathology were selected. Two review authors independently selected studies for inclusion, extracted data and assessed risk of bias. Pooled summary estimates of sensitivity and specificity with 95% CI were calculated.

*Results*  Seven studies were included, all describing the diagnostic accuracy of CT in identifying lumbar disc herniation. Six studies used surgical findings as the reference standard and were considered sufficiently homogenous to carry out a meta-analysis. The pooled summary estimate of sensitivity was 77.4% and specificity was 73.7%.

*Conclusions*  We found no studies evaluating the accuracy of CT for pathologies such as vertebral cancer, infection and fractures and this remains unclear. Our results should be interpreted with some caution. Sensitivity and specificity, regarding the detection of lumbar disc herniation, showed that a substantial part of the patients is still classified as false-negative and false-positive. In future, the diagnostic performance of CT must be assessed in high quality prospective cohort studies with an unselected population of patients with low back pain.

**Keywords**  Computed tomography · Low back pain · Diagnostic imaging · Systematic review

## Introduction

Low back pain (LBP) is usually defined as pain, muscle tension, or stiffness localised below the costal margin and above the inferior gluteal folds, with or without sciatica [1]. With a lifetime prevalence from 11 to 84% LBP is a major health problem worldwide, and causes a substantial economic burden in Western societies [2–7].

LBP is occasionally the presenting symptom of an underlying pathology such as radiculopathy or spinal stenosis or another specific spinal condition [8]. The diagnostic process is to distinguish 'simple' back pain from back pain due to serious underlying diseases or neurologic impairments [9]. Guidelines recommend starting the diagnostic triage with history taking and physical examination, in order to identify 'red-flags' and classify patients into one

R. M. van Rijn · A. P. Verhagen (✉) · B. W. Koes
Department of General Practice, Erasmus MC,
University Medical Centre, P.O. Box 2040,
3000 CA Rotterdam, The Netherlands
e-mail: a.verhagen@erasmusmc.nl

M. Wassenaar
Department of Clinical Neurophysiology,
University Medical Centre Utrecht, Utrecht, The Netherlands

R. W. J. G. Ostelo · M. R. de Boer · M. W. van Tulder
Department of Health Sciences and EMGO-Institute for Health
and Care Research, Faculty of Earth and Life Sciences,
VU University Amsterdam, Amsterdam, The Netherlands

A. Z. Ginai
Department of Radiology, Erasmus MC,
University Medical Centre, Rotterdam, The Netherlands

of three categories: serious spinal pathology, nerve root pain/radicular pain, and nonspecific LBP [8, 10].

Routine lumbar-spine imaging is not recommended in patients with LBP without symptoms suggesting serious underlying conditions [8, 10, 11]. However, if serious spinal pathology is suspected based on red-flags, diagnostic imaging could be performed, since delayed diagnosis and treatment are associated with poorer outcomes [8]. One of the diagnostic imaging techniques available for this purpose is computed tomography (CT). CT nowadays plays a vital role in spinal imaging and has largely replaced invasive imaging techniques, such as myelography, epidural venography and epidurography, particularly because CT is associated with less morbidity than invasive techniques [9, 12]. Caution is, however, necessary in the choice of CT as an imaging modality particularly in younger patients due to gonadal radiation dose particularly with repeated examinations. For this reason, in many clinical guidelines magnetic resonance imaging (MRI) is the imaging modality of choice. CT is suggested to be the primary imaging technique to depict disorders of bone structures [15]. CT is also used for detection of chronic morphologic changes and has a well-recognized role in the diagnosis of spinal stenosis, herniated nucleus pulposus and facet joint abnormalities [13, 14]. Additionally, compared to MRI, CT is cheaper, the total testing time is shorter, and the availability of CT scanners is larger in hospital settings. The disadvantages of CT, however, is the radiation dose particularly cumulative dose with repeat examinations in younger patients. Moreover, even when MRI is readily available, the need for a CT of the lumbar spine in the presence of a cardiac pacemaker seems to be increasing with an aging population.

Estimates of the diagnostic accuracy of CT scans vary considerably across primary diagnostic studies. Potential sources of heterogeneity include: difference in considered pathology, variation in CT protocols, differences in study design, included study populations, or the methodological quality of the studies. Therefore, our aim is to provide evidence on the diagnostic accuracy of CT in patients with LBP or sciatica with symptoms suspected to be caused by specific underlying pathology. Sciatica is here defined as nerve root pain or radiating leg pain. We also aim to assess the potential influence of various sources of heterogeneity on the outcomes.

## Methods

### Design

Systematic review of diagnostic accuracy studies.

### Search strategy

We systematically searched Medline, Embase and CINAHL databases (until December 2009). The search strategy we used was developed to identify publications for four separated systematic reviews. These systematic reviews all concern the diagnostic test accuracy of imaging techniques (MRI, CT, X-ray, or myelography) for identifying or excluding lumbar spinal pathology.
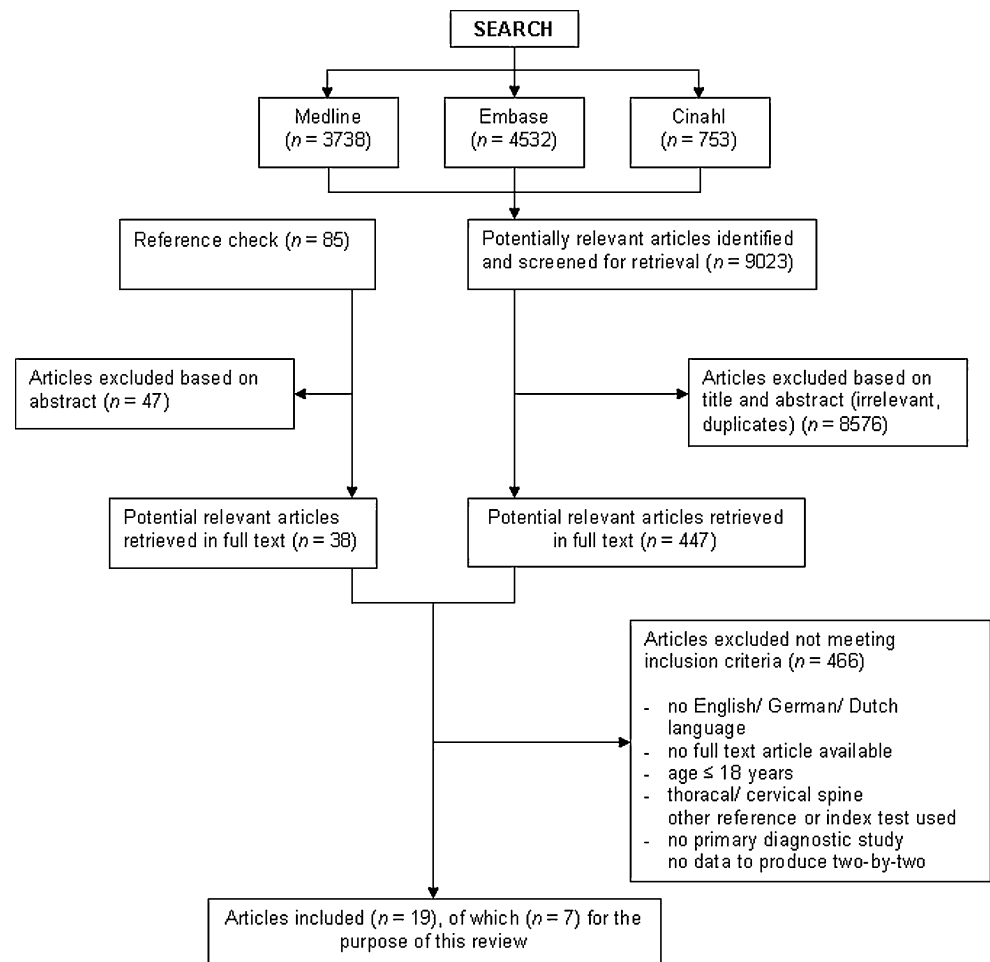
### Study selection

Two review authors (AV, MW) independently selected the articles, based on title and abstract (Fig. 1). For final inclusion the studies had to fulfill the following criteria: (1) the diagnostic accuracy of CT was assessed in adult patients with LBP suspected to be caused by specific pathology (i.e. radicular syndrome, spinal stenosis, spinal tumors, spinal fractures, spinal infection/inflammation, spondylolisthesis, spondylolysis, ankylosing spondylitis, disc displacement, osteoporotic fractures, and other degenerative disc diseases), (2) the results were compared with those of a reference test (i.e. findings at surgery, expert panel opinion, diagnostic work-up, or MRI), (3) the design was a case–control or cohort study; either prospective or retrospective, and (4) the results were published as full reports with sufficient data to construct diagnostic two-by-two tables. Disagreements were resolved by consensus; a third review author (MvT) was consulted in case of persisting disagreement.

### Data extraction and risk of bias assessment

Data extraction was performed by two review authors (RvR, MW) independently using a standardised form. Data were extracted on: (1) study design; prospective or retrospective observational study, (2) characteristics of study population; setting, age, gender, pathology considered, duration and history of LBP, inclusion and exclusion criteria, enrollment, number of subjects (enrolled, eligible), level of measurement (3) test characteristics; type of index test, type of reference test, year and methods of execution, outcome scales, and (4) diagnostic parameters; two-by-two table, or, if not available, relevant parameters to reconstruct this table.

Two independent review authors (MW, RvR) assessed the risk of bias of each included study using the QUality Assessment of Diagnostic Accuracy Studies (QUADAS) tool [16, 17]. The QUADAS tool consisted of 11 items that referred to internal validity. In addition, we identified nine additional items described in the Cochrane Handbook for Diagnostic Test Accuracy Reviews [17]. These additional items were of relevance to this review and were also

**Fig. 1** Flow chart of selected articles



```
                              ┌──────────────┐
                              │   SEARCH     │
                              └──────────────┘
        ┌──────────────┬──────────────┬──────────────┐
  ┌───────────┐   ┌───────────┐   ┌───────────┐
  │  Medline  │   │  Embase   │   │   Cinahl  │
  │ (n = 3738)│   │ (n = 4532)│   │ (n = 753) │
  └───────────┘   └───────────┘   └───────────┘
```

Reference check (n = 85)

Potentially relevant articles identified and screened for retrieval (n = 9023)

Articles excluded based on abstract (n = 47)

Articles excluded based on title and abstract (irrelevant, duplicates) (n = 8576)

Potential relevant articles retrieved in full text (n = 38)

Potential relevant articles retrieved in full text (n = 447)

Articles excluded not meeting inclusion criteria (n = 466)

- no English/ German/ Dutch language
- no full text article available
- age ≤ 18 years
- thoracal/ cervical spine other reference or index test used
- no primary diagnostic study no data to produce two-by-two

Articles included (n = 19), of which (n = 7) for the purpose of this review

scored. The 20 items were scored as "yes", "no", or "unclear" according to the classification definition described in Appendix 1. A radiologist (AG) was consulted for the assessment of the used technology (item 13). Disagreements were resolved by consensus. In case of persisting disagreement a third review author (AV) was consulted. We did not apply weights to the different items and did not use a summary score since the interpretation of summary scores was problematic and potentially misleading [18, 19].

Data synthesis and analysis

From each included study we used the two-by-two table to calculate sensitivity and specificity with the corresponding 95% confidence intervals (95% CI). For a descriptive analysis, sensitivity and specificity were presented in forest plots. Besides, we plotted the results on a receiver operating curve (ROC) plot of sensitivity against 1-specificity.

For meta-analysis of pairs of sensitivity and specificity we used a bivariate random effects method [20]. This method provides summary estimates of sensitivity and specificity with corresponding 95% CI while dealing with sources of variation within and between studies and any correlation that might exist between sensitivity and specificity. We calculated a 95% confidence ellipse around the summary estimate of sensitivity and specificity, and plotted the results in ROC space. We only conducted a meta-analysis if studies show sufficient homogeneity (i.e. same pathology, same reference standard, comparable population, same study design). Analyses were carried out using STATA 10 software. All findings were presented in a summary of results Table 2, which included summary estimates of sensitivity and specificity, prior probabilities, diagnostic odds ratio, and likelihood ratios for the diagnostic accuracy of CT.

Several factors may contribute to heterogeneity in diagnostic performance across studies. We investigated the potential influence of differences in pathology, and reference standards used in the primary studies on sensitivity and specificity, both by comparing subgroups. If sufficient data were available, we assessed the possible bias introduced by negative scores on several important items on the risk of bias assessment. These items were independently

added as a covariate to the bivariate model. The results were presented graphically and in a summary of results table.

## Results

### Literature search

Our search resulted in 9023 potentially relevant articles of which 447 were retrieved in full text. The additional search and reference check resulted in 85 possible relevant articles, of which 38 were retrieved in full text. Finally, 19 articles met our inclusion criteria and were eligible for at least one of the four separate reviews conducted on the diagnostic accuracy of imaging in adult LBP patients to identify or exclude specific pathology (Fig. 1). Of these, seven articles focused on CT and were included in this review [21–27]. All studies described the diagnostic accuracy of CT in identifying lumbar disc herniation (Table 1).

### Risk of bias assessment

Figure 2 presents the results of the individual studies. The initial agreement of the reviewers was 78% (109 of 140 items). The disagreements were resolved by consensus. All studies used an acceptable reference standard, avoided differential verification, and pre-specified their objectives (items 2, 5 and 19). None of the studies reported enough information to assess the items on the delay between index test and reference test, observer variation, instrument variation, appropriate patient subgroups, appropriate sample size, and whether treatment or intervention was initiated between index test and reference test (items 3, 15, 16, 17, 18, and 20). The majority of studies poorly described the selection of patients, blinding of reference test results, and whether cut-off values were pre-specified (item 1, 7 and 12), resulting in a high risk of selection and reviewer bias. In two studies [22, 26] not all patients received confirmation of their diagnosis by a reference test (item 4), and in four studies [21, 23, 25, 26] the technology of CT used was changed since the study was carried out (item 13). Since these two items were thought to influence the reported sensitivity and specificity we added them individually as covariates to the bivariate analysis.

### Findings

For each study the extracted data (2 × 2 table) and sensitivity and specificity are presented in a forest plot (Fig. 3).

All studies described the accuracy of CT in identifying lumbar disc herniation, containing a total of 498 discs

explorations and 296 measurements on patient level. The prior probability of lumbar disc herniation, varied from 49.2 [24] to 90.5% [21]. In these studies, lumbar disc herniation was defined as herniated nucleus pulposus, including protruded, extruded or sequestrated disc, or causing nerve root compression. One study used expert panel consensus, a four stage process, as the reference standard resulting in a sensitivity of 94% (95% CI 73–100%) and a specificity of 64% (95% CI 35–87%) [27]. Six studies used surgical findings as the reference standard [21–26]. We considered these studies sufficiently homogenous for a meta-analysis. The sensitivity and specificity of CT in identifying lumbar disc herniation in these studies ranged from 59 to 92% and from 45 to 87%, respectively. The results of the bivariate analysis are presented in the Table 2 and plotted in a ROC space (Fig. 4). The pooled summary estimate of sensitivity was 77.4% (95% CI 66.2–85.7%) and the pooled summary estimate of specificity was 73.7% (95% CI 61.8–82.9%), resulting in a positive likelihood ratio of 2.94, a negative likelihood ratio of 0.31, and a diagnostic odds ratio of 9.61.

The influence of pre-defined potential sources of heterogeneity was determined by adding each individual QUADAS item as covariate to the bivariate model (Table 2). We assessed the influence of partial verification bias and used CT technology (items 4 and 13). Adding the item on partial verification bias to the model resulted in a pooled summary estimate of sensitivity and specificity of 76.7% (95% CI 64.7–85.6%) and 73.4% (95% CI 61.2–82.8%), respectively. Summary estimates of sensitivity and specificity changed to 79.1% (95% CI 65.0–88.5%) and 76.0% (95% CI 60.1–87.0%), respectively after adding the item on use of an appropriate CT technique as covariate to the model. The item for selection bias (item 1) was poorly described and could, therefore, not be added as a covariate to the model.

We were unable to evaluate the influence of differences in pathology and different reference standards on sensitivity and specificity, since six out of seven studies investigated the accuracy of CT in identifying lumbar disc herniation with surgical findings as the reference standard. Exploratory analysis on the influence of the use of a prospective versus a retrospective design and measurements at disc level versus patient level did not resulted in a different accuracy of CT.

## Discussion

This review included seven studies on lumbar disc herniation, and found a pooled summary estimate of sensitivity of 77.4% and specificity of 73.7% for CT compared to surgical findings. This means that, a substantial part of the

**Table 1** Characteristics of included studies

| Author | Design and setting | Patients | Target condition (prevalence) | Level of measurement | Index test | Reference test |
|---|---|---|---|---|---|---|
| Firooznia et al. [21] | Prospective. Secondary care, Germany | 100 patients who underwent surgery for sciatica, and had CT of lumbar spine before surgery: 61% male, mean age 49 (19–76) years | Disc prolapse (90.5%) | Disc level; 116 levels assessed of 100 patients | CT: GE 8800 CT/T, 25 cm circular calibration, 250–400 mA, 120 kV, 9.6 s speed, 5 mm slice thickness, with a radiation to the patient per slice of 2.5–4.2 rad | Surgical findings |
| Forristall et al. [22] | Prospective. Secondary care, USA | 25 patients of which clinical findings were consistent with a HNP documented by positive findings on MRI or contrast CT: 78% male, mean age 45 (22–74) years | HNP with neural compression (77.4%) | Disc level; 31 levels assessed of 25 patients | CT: Picker 1200 Synerview, 14 cm, 65 mA, 130 kV, 5 mm slice thickness, 5 ml of Amipaque 180 mg I/ml | Surgical findings |
| Jackson et al. [23] | Prospective. Secondary care, USA | 124 patients with LBP and leg pain due to degenerative spinal pathology refractory to conservative management: 70% male, mean age 43 (21–76) years | HNP: protruded, extruded, and sequestrated disc (54.1%) | Disc level; 231 levels assessed of 124 patients | CT: Siemens Somatom, 5 mm slice thickness with 1 mm overlap using bone and soft tissue settings | Surgical findings |
| Jackson et al. [24] | Prospective. Secondary care, USA | 59 patients with LBP and leg pain due to degenerative spinal pathology refractory to conservative management: 56% male, mean age 40 (18–70) years | HNP: protruded, extruded, and sequestrated disc (49.2%) | Disc level; 120 levels assessed of 59 patients | CT: Siemens Somatom, 5 mm slice thickness with 1 mm overlap using bone and soft tissue settings | Surgical findings |
| Schaub et al. [25] | Retrospective. Secondary care, Swiss | 29 patients with recurring symptoms after lumbar disk surgery: 48% male, mean age 49 (SD: 13) years | HNP (62.1%) | Patient level | CT | Surgical findings |
| Schipper et al. [26] | Prospective. Secondary care, Netherlands | 235 patients with radiating leg pain, with or without back pain, with feelings of numbness, or with paresis and referred to the neurosurgical department: 61% male, mean age 43 years | HNP: an asymmetric protruding disk, obliteration of the epidural fat, compression or displacement of the nerve root, indentation of the dural sac (83.8%) | Patient level | CT: Philips Tomoscan 350, 200 As, 120 kV, 3 mm slice thickness | Surgical findings |
| Thornbury et al. [27] | Prospective. Secondary care, USA | 32 patients with acute LBP and radicular pain in whom the diagnosis of HNPNC was sufficiently probable | HNP with nerve root compression (56.3%) | Patient level | CT: Siemens Somatome, 4 mm slice thickness, 125 kV, 550 mA s | Expert panel: four stages; review clinical material, information of medical record and follow up survey including details of therapy, blinded reading of results of one of the two radiologic examinations, blinded reading of other radiologic examination |

**Fig. 2** Risk of bias scores for
each included study
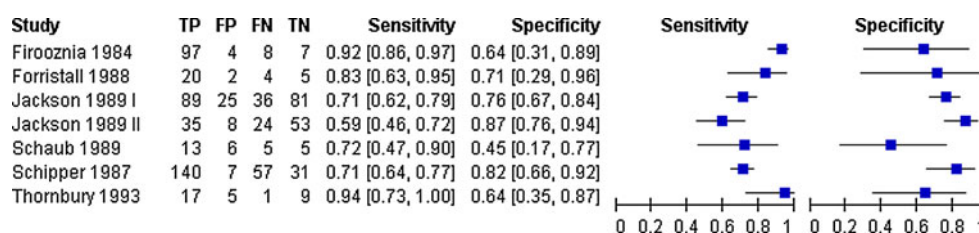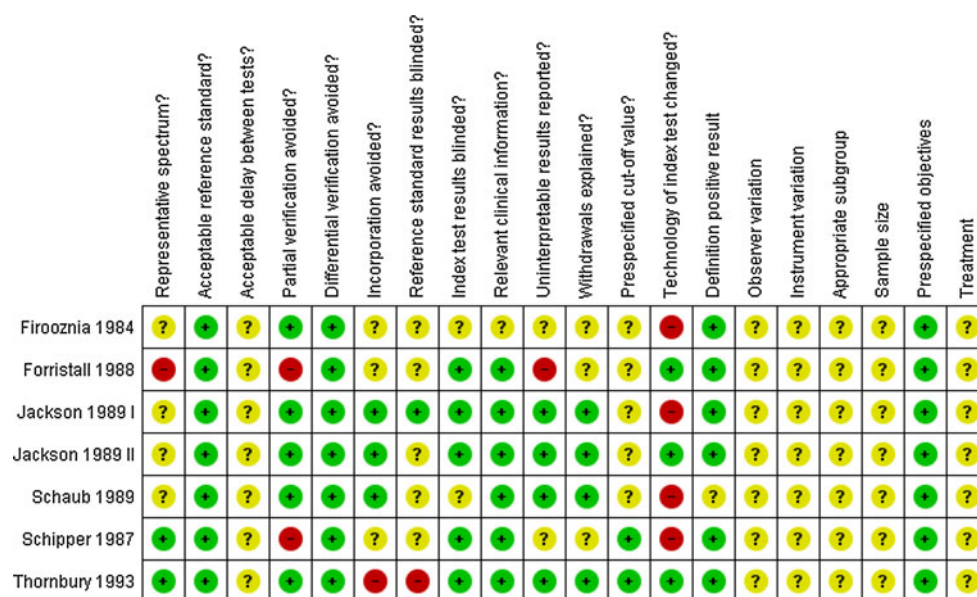


**Fig. 3** Forest plot of seven comparisons of the seven included studies describing lumbar disc herniation as specific pathology with the estimated sensitivity and specificity with accompanying 95% confidence intervals. *TP* true-positive, *FP* false-positive, *FN* false-negative, *TN* true-negative

**Table 2** Results of the bivariate analysis with summary estimates of sensitivity, specificity, positive likelihood ratio (LR+), and negative likelihood ratio (LR−) and the accompanying diagnostic odds ratio (DOR) and the prior probability of lumbar disc herniation

| Target condition | Reference test | Covariates | Sensitivity (95% CI) | Specificity (95% CI) | Mean prior probability (range) | LR+ (95% CI) | LR− (95% CI) | DOR (95% CI) |
|---|---|---|---|---|---|---|---|---|
| Lumbar disc herniation | Surgery | | 0.77 (0.66–0.86) | 0.74 (0.62–0.83) | 69.5% (49.2–90.5%) | 2.94 (2.12–4.09) | 0.31 (0.22–0.43) | 9.61 (6.22–14.84) |
| | | QUADAS item 4: partial verification | 0.77 (0.65–0.86) | 0.73 (0.61–0.83) | | 2.88 (2.07–4.00) | 0.32 (0.22–0.46) | 9.08 (5.58–14.77) |
| | | QUADAS item 13: used technology | 0.79 (0.65–0.89) | 0.76 (0.60–0.87) | | 3.30 (1.79–6.07) | 0.27 (0.15–0.50) | 12.01 (4.22–34.17) |

patients is still classified as false-negative and false-positive. The use of newer CT technique resulted in a slightly better accuracy compared to the use of old CT technology.

The results of this review should be interpreted with caution. First, prior probabilities of the underlying pathologies of LBP showed a large variation. The diagnostic value of CT also depends both on the prior probability of the underlying pathology in the investigated population. In general, a high prior probability results in a high positive diagnostic value and a low negative diagnostic value, and vice versa [28]. The large variation of prior probabilities might be due to the selection procedure of the patients, as in five out of seven studies the selection procedure was unclear or inadequate and therefore selection bias might have occurred. Besides, all included studies were performed in a secondary care setting, where
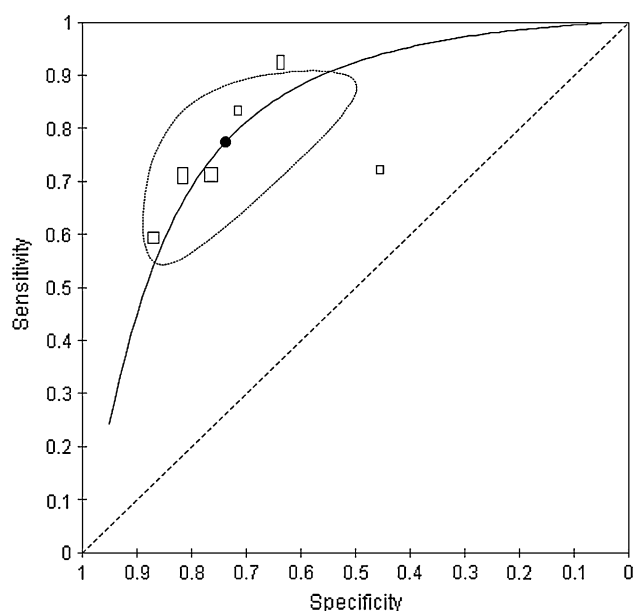
**Fig. 4** Summary ROC plots of sensitivity and specificity of six studies describing the diagnostic accuracy of computed tomography with surgical findings as the reference standard and lumbar disc herniation as specific pathology. The width of the rectangles is proportional to the number of patients with possible or without lumbar disc herniation; the height of the blocks is proportional to the number of patients with lumbar disc herniation (proven or probable). The *solid line* is the summary ROC curve; the *black spot* is the mean value for sensitivity and specificity; the *ellipse around the black spot* represent the 95% confidence interval around this summary estimate

patients often will have a higher prior probability due to referral.

Secondly, because of the absence of a clear gold reference standard studies were included if findings at surgery, expert panel opinion, diagnostic work-up, or MRI was used as reference standard. Finally, one study used expert panel consensus and six studies surgical findings as the reference standard. The problem with surgical populations is that only patients with a strong suspicion on a specific underlying pathology are subjected to surgery. Therefore, the results of these studies can easily be biased, leading to an overestimation of the diagnostic accuracy of the index test.

Thirdly, the accuracy of an index test also depends on the reliability of the test, definition of a positive result and used technology. As CT requires some degree of expertise it is not surprisingly that the reliability of CT varies considerably. None of the studies reported data on the observer variation. Therefore, the extent of the effect on the results cannot be estimated. Also, the CT technology used can be of influence on the diagnostic accuracy. Assessing the effect of the use of CT technology resulted in an increase in the sensitivity and specificity when using a newer CT technology. Most CT technology used in the studies found are rather outdated as the most recent study is published in

1993, probably modern technology, not yet evaluated in the studies available will show better results.

Finally, the diagnostic imaging studies reported their results on patient level as well as on disc level. Presenting the results on disc level will lead to multiple inclusions of the same patients. Besides, patients with signs of lumbar disc herniation are more likely to be subjected to multiple level testing than patients without these signs. This might lead to an overestimation of the diagnostic performance of CT. Here, four studies presented their results on disc level only, but an explorative subgroup analysis did not result in different pooled summary estimates.

Strengths and weakness of the review

This is the first systematic review that provides evidence on the diagnostic accuracy of CT in LBP patients. One of the limitations of this study was the use of a filter to limit the primary literature search. The filter was targeted on study design to overcome indexing problems related to terms like sensitivity, specificity, accuracy or predictive value. After a random check we assumed using this filter would not lead to missing relevant studies. Second, the generalisability of the results is limited mainly by poor reporting in the original studies, which lead to many unclear or inadequate scores on several QUADAS items. This means that the potential influence of bias is difficult to assess [29].

Implication to clinical practice

The summary estimates for sensitivity and specificity for CT in identifying lumbar disc herniation may be acceptable, but also demonstrates that a substantial part of the patients will be wrongly diagnosed. However, the accuracy of CT might differ between pathology, but no studies were found evaluating the accuracy of CT for pathologies such as vertebral cancer, infection and fractures and this remains unclear. The applicability to clinical practice also depends on the role to which the diagnostic test is allocated [30]. Here most studies present the separate diagnostic value of CT, although in clinical practice CT is part of the diagnostic process which might lead to a better diagnostic performance as a whole. Therefore, more research is needed before our results can be translated to clinical practice and policy.

Implication for research

Given the possible advantages of CT over MRI future research should focus on the diagnostic performance of up to date CT technology assessed in high quality prospective cohort studies with an unselected population of patients with LBP. In order to provide clear evidence when to use

CT or not, analyses should be done on patient level and in combination with other diagnostic tools. Furthermore, in order to improve accuracy and completeness of reporting of accuracy studies, future studies should comply with the STARD initiative [31].

## Appendix 1: Items for Quality Assessment of Diagnostic Accuracy Studies (QUADAS) with 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 or patients with other causes of LBP than the cause aimed to identify. Score also 'no' if non-response is high and selective, or there is clear evidence of selective sampling.

**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 specific lumbar-spine pathology 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. Since there is no 'gold reference standard' available for diagnosing spinal pathology we accepted the following as reference standards: (1) findings at surgery, (2) clinical opinion/expert panel opinion, (3) diagnostic work-up, (4) MRI.

**Classify as 'yes'** if one of these procedures is used as reference standards.

**Classify as 'no'** if you seriously question the methods used or other procedure/test is used as reference standard.

**Classify as 'unclear'** if insufficient information is given on the reference standard.

### 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 index test and the reference standard is 1 week or less.

**Classify as 'no'** if the time period between index test and the reference standard is longer than 1 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.

### 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 computed tomography, whereas others receive only diagnosis by expert opinion, diagnostic work-up or other imaging tests. 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 diagnostic imaging 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 expert opinion or diagnostic work-up results only (regardless of knowledge of the results of diagnostic imaging tests), the index test is *not* part of the reference standard. However, if the final diagnosis is based on the results of both expert opinion or diagnostic work-up *and* a positive computed tomography 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 reference standard results interpreted without knowledge of the results of the index test?*

Interpretation of the results of computed tomography 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 reference test precedes the index test, which is unlikely in case of surgery, interpretation of the results of the reference test will usually be without knowledge of the results of the index test.

**Classify as 'yes'** if the test results of the reference standard are interpreted blind to the results of the index 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 index test.

**Classify as 'unclear'** if insufficient information is given on independent or blind assessment of the index test or reference standard.

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

Interpretation of the results of computed tomography 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.

**Classify as 'yes'** if the test results of computed tomography are interpreted blind to the results of the reference 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 reference 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 clinical data, such as age, gender, symptoms, history of LBP, previous treatments or other test results 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 diagnostic imaging 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 would normally be available when the test results are interpreted and similar data are available in the study.

**Classify as 'no'** if this is not the case, e.g. if other test results are available that cannot 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.

*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. Also 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). Also 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 will be scored under item 1.

### Additional QUADAS items

### 12. If a cut-off value has been used, was it established before the study was started (pre-specified cut-off value)?

**Classify as 'yes'** if chosen cut-off values are mentioned in "Methods" section as chosen at forehand, preferably with reason for that specific choice.

**Classify as 'no'** if cut-off values are mentioned in "Result"/"Discussion" section and they are assumed not to be chosen at forehand.

**Classify as 'unclear'** if insufficient information is given on which cut-off values have been used.

### 13. Is the technology of the index test likely to have changed since the study was carried out?

**Classify as 'yes'** if the technology of the index test is not likely to have changed since the study was carried out. Most recent technology was used.

**Classify as 'no'** if the technology of the index test is likely to have changed since the study was carried out due to development over time.

**Classify as 'unclear'** if insufficient information is given on the technology used for the index test.

### 14. Did the study provide a clear definition of what was considered to be a "positive" result?

The index test can be prone to subjective judgements with several possible definitions of what is considered a positive test result. Furthermore, several methods of performing the computed tomography test have been described, and several cut-offs have been proposed. Consequently, it is essential that an adequate description is given of the methods that have been used to define a positive or negative test result.

**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.

### 15. Were data on observer variation reported? and within acceptable range?

Studies on the reproducibility of computed tomography in LBP patients 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 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.

### 16. Were data on instrument variation reported?

This item was interpreted and assessed as reporting of measurement variation.

**Classify as 'yes'** if the paper provides information on instrument variation showing no or acceptable measurement variation.

**Classify as 'no'** if information is provided on instrument variation demonstrating variation in measurement.

**Classify as 'unclear'** if the paper provides no information to assess instrument variation of the used tests.

### 17. Were data presented for appropriate patient subgroups?

**Classify as 'yes'** if subgroups were considered beforehand (mentioned already in "Methods" section) and data of subgroups were adequately presented, or if subgroups were not considered, and it seems logic from the context that stratification in subgroups was not required.

**Classify as 'no'** if data on subgroups were presented but assumed not to be chosen beforehand (not mentioned in the "Methods" section). Or if subgroups were not considered, although it seems logic from the context that stratification in subgroups was required.

**Classify as 'unclear'** if insufficient information was given for considerations on the use of a possible subgroup analysis.

### 18. Was an appropriate sample size included?

**Classify as 'yes'** if a sample size calculation was performed in advance, and the same sample size has been included, or if not calculated but if the paper provides information on reasons for a chosen specific sample size which seems appropriate.

**Classify as 'no'** if no sample size calculation was performed in advance, or the number of patients calculated was not included in the study. Or if no other reasons for the chosen sample size were given.

**Classify as 'unclear'** if insufficient information was given on considerations of a sample size.

### 19. Were objectives pre-specified?

Studies should be designed as primary diagnostic studies in which the objectives reflects assessment of the diagnostic accuracy of the index test.

**Classify as 'yes'** if the objectives were mentioned in the "Introduction" section (as pre-specified), and the data presented adequately reflects these objectives.

**Classify as 'no'** if the objectives were mentioned in the "Introduction" section (as pre-specified), however, the data presented did not reflect these objectives, i.e. the real objectives were clarified or adjusted in the "Result"/ "Discussion" section. Or the objectives were specified according to the data or results.

**Classify as 'unclear'** if the paper does not provide information on the pre-specification of the objectives, i.e. objectives were not stated in the "Introduction" section.

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

If index tests and reference standard are not performed within a short amount of time, some type of treatment or 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 or intervention is given in the time period between index test and reference standard.

**Classify as 'no'** if a treatment or intervention is given before both index test and reference standard were performed that could possibly influence the prognosis of LBP and its outcome on index or reference test.

**Classify as 'unclear'** if there is insufficient information regarding treatment or intervention performed before the index and reference test were both performed.

## References

1. Koes BW, van Tulder MW, Thomas S (2006) Diagnosis and treatment of low back pain. BMJ 332:1430–1434
2. Ekman M, Johnell O, Lidgren L (2005) The economic cost of low back pain in Sweden in 2001. Acta Orthop 76:275–284
3. Maniadakis N, Gray A (2000) The economic burden of back pain in the UK. Pain 84:95–103
4. Rizzo JA, Abbott TA, Berger ML (1998) The labor productivity effects of chronic backache in the United States. Med Care 36:1471–1488
5. van Tulder MW, Koes BW, Bouter LM (1995) A cost-of-illness study of back pain in The Netherlands. Pain 62:233–240
6. Wenig CM, Schmidt CO, Kohlmann T, Schweikert B (2009) Costs of back pain in Germany. Eur J Pain 13:280–286
7. Walker BF (2000) The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. J Spinal Disord 13:205–217
8. Chou R, Qaseem A, Snow V, Casey D, Cross JT Jr, Shekelle P et al (2007) Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med 147:478–491
9. Jarvik JG, Deyo RA (2002) Diagnostic evaluation of low back pain with emphasis on imaging. Ann Intern Med 137:586–597
10. van Tulder M, Becker A, Bekkering T, Breen A, del Real MT, Hutchinson A et al (2006) Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care. Eur Spine J 15(Suppl 2):S169–S191
11. Chou R, Fu R, Carrino JA, Deyo RA (2009) Imaging strategies for low-back pain: systematic review and meta-analysis. Lancet 373:463–472
12. Finch P (2006) Technology Insight: imaging of low back pain. Nat Clin Pract Rheumatol 2:554–561
13. Hall FM (1980) Back pain and the radiologist. Radiology 137:861–863
14. Lurie JD (2005) What diagnostic tests are useful for low back pain? Best Pract Res Clin Rheumatol 19:557–575
15. Verhaar JAN, van Mourik JBA (2008) Orthopedie. Bohn Stafleu van Loghum, Houten
16. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J (2004) Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess 8:iii, 1–234
17. Smidt N, Deeks J, Moore T (2005) Cochrane handbook for cochrane reviews of diagnostic test accuracy. The Cochrane Collaboration, UK
18. Juni P, Witschi A, Bloch R, Egger M (1999) The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282:1054–1060
19. 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
20. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 58:982–990
21. Firooznia H, Benjamin V, Kricheff II, Rafii M, Golimbu C (1984) CT of lumbar spine disk herniation: correlation with surgical findings. AJR Am J Roentgenol 142:587–592
22. Forristall RM, Marsh HO, Pay NT (1988) Magnetic resonance imaging and contrast CT of the lumbar spine. Comparison of diagnostic methods and correlation with surgical findings. Spine (Phila Pa 1976) 13:1049–1054

23. Jackson RP, Becker GJ, Jacobs RR, Montesano PX, Cooper BR, McManus GE (1989) The neuroradiographic diagnosis of lumbar herniated nucleus pulposus: I. A comparison of computed tomography (CT), myelography, CT-myelography, discography, and CT-discography. Spine (Phila Pa 1976) 14:1356–1361

24. Jackson RP, Cain JE Jr, Jacobs RR, Cooper BR, McManus GE (1989) The neuroradiographic diagnosis of lumbar herniated nucleus pulposus: II. A comparison of computed tomography (CT), myelography, CT-myelography, and magnetic resonance imaging. Spine (Phila Pa 1976) 14:1362–1367

25. Schaub B, Gratzl O, Bahous I (1989) The value of myelography, computerized tomography and course of pain for the diagnosis of recurrent herniated disk. Schweiz Med Wochenschr 119:1185–1190

26. Schipper J, Kardaun JW, Braakman R, van Dongen KJ, Blaauw G (1987) Lumbar disk herniation: diagnosis with CT or myelography. Radiology 165:227–231

27. Thornbury JR, Fryback DG, Turski PA, Javid MJ, McDonald JV, Beinlich BR et al (1993) Disk-caused nerve compression in patients with acute low-back pain: diagnosis with MR, CT myelography, and plain CT. Radiology 186:731–738

28. Gordis L (2009) Epidemiology. Saunders Elsevier, Philadelphia

29. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al (1999) Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 282:1061–1066

30. Bossuyt PM, Irwig L, Craig J, Glasziou P (2006) Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 332:1089–1092

31. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 138:40–44