Key issues in diagnosis and targeted treatment of low back pain

Christopher Seungmin Han

B App Sci (OT), DPT

A thesis submitted to fulfil requirements for the degree of

Doctor of Philosophy

School of Public Health, Faculty of Medicine, and Health

The University of Sydney

2023

Supervisors statement

As supervisors of Christopher Seungmin Han's doctoral work, we certify that we consider the thesis 'Key issues in diagnosis and targeted treatment of low back pain' to be suitable for examination.

Professor Christopher Maher

Sydney Musculoskeletal Health

School of Public Health

Faculty of Medicine and Health

The University of Sydney

Date: 13th December 2023

Professor Mark Hancock

Department of Health Sciences

Faculty of Medicine, Health and Human Sciences

Macquarie University

Date: 13th December 2023

Candidate's statement

This thesis is submitted to The University of Sydney in fulfilment of the requirements for the Degree of Doctor of Philosophy.

I, Christopher Seungmin Han, hereby declare that this manuscript is my own work and that it contains no material previously published or written by another person except where acknowledged in the text. Nor does it contain material which has been submitted, either in full or in part, for a degree at this or any other institution.

I, Christopher Seungmin Han, understand that if I am awarded a higher degree for my thesis entitled 'Understanding the role of imaging in low back pain' being lodged herewith for examination, the thesis will be lodged in the University Library and be available immediately for use. I agree that the University Librarian (or in the case of a department, the Head of Department) may supply a photocopy or microform of the thesis to an individual for research or study or to a library.

_____ Date: 13th December 2023

Authorship Attribution Statement

Chapter 2 of this thesis is published as:

Han CS, Hancock MJ, Maher CG. Reconsidering non-specific low back pain: where to from here? Spine J. 2022 Dec;22(12):1927-1930. doi: 10.1016/j.spinee.2022.08.001. Epub 2022 Aug 6. PMID: 35944828.

I was responsible for the Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

Chapter 3 of this thesis is published as:

Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Maher CG. Red flags to screen for vertebral fracture in patients presenting with low back pain. Cochrane Database Syst Rev. 2022 Jul 20;2022(7):CD014461. doi: 10.1002/14651858.CD014461. PMCID: PMC9298669.

I was responsible for the Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

Chapter 4 of this thesis is published as:

Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Chen Q, Maher CG. Red flags to screen for vertebral fracture in people presenting with low back pain. Cochrane Database Syst Rev. 2023 Aug 24;8(8):CD014461. doi: 10.1002/14651858.CD014461.pub2. PMID: 37615643; PMCID: PMC10448864.

I was responsible for the Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

Chapter 5 of this thesis is published as:

Han CS, Hancock MJ, Sharma S, Sharma S, Harris IA, Cohen SP, Magnussen J, Maher CG, Traeger AC. Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review. EClinicalMedicine. 2023 Apr 6;59:101960. doi: 10.1016/j.eclinm.2023.101960. PMID: 37096189; PMCID: PMC10121397.

I was responsible for the Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

Chapter 6 of this thesis is published as:

Han CS, Maher CG, Steffens D, Diwan A, Magnussen J, Hancock EC, Hancock MJ. Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review. J Physiother. 2023 Apr;69(2):79-92. doi: 10.1016/j.jphys.2023.02.007. Epub 2023 Mar 11. PMID: 36914521.

I was responsible for the Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

Acknowledgements

My PhD journey was challenging in many respects, yet very rewarding. There were many life milestones accomplished throughout my candidature, and there are many people who need to be thanked.

Firstly, I would like to give my deepest thanks and gratitude to my incredible supervisors Chris Maher and Mark Hancock. Chris, thank you for being the unselfish leader that you are, sharing your incredible knowledge and skills without hesitation, and always being so present and supportive during my candidature. Mark, thank you for creating such an inspiring learning environment from my first encounter with you as a Physiotherapy student at Macquarie University to now completing my PhD, thank you for sharing your passion for our profession, and your endearing advice on career, family, and life. In the past three years, I have learned from both my supervisors to be critical in my thinking and craftsmanship, the importance of balancing between being a clinician and a researcher, and the continual reminder that we owe it to our patients to strive for excellence.

I would like to express my deepest gratitude to the wonderful group of people at the Institute for Musculoskeletal Health that have been part of my journey. I would also like to thank my Physiotherapy colleagues for their support and providing opportunities to still excel my career whilst I completed my PhD.

Lastly, I owe an incredible thank you to my family and dearest friends for their unconditional love and support, and always encouraging personal growth. To my family, what you have

accomplished in your lives and continue to accomplish is my greatest aspiration. To my wife, there are no words to express my gratitude for your continual support in everything that I do, and for doing it with so much love.

Table of contents

Publications and presentations

Parts of the work presented in this thesis have been published and/or presented in the following forms:

Publications in peer reviewed journals

1. Han CS, Hancock MJ, Maher CG. Reconsidering non-specific low back pain: where to from here? Spine J. 2022 Dec;22(12):1927-1930. doi: 10.1016/j.spinee.2022.08.001. Epub 2022 Aug 6. PMID: 35944828.
2. Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Maher CG. Red flags to screen for vertebral fracture in patients presenting with low back pain. Cochrane Database Syst Rev. 2022 Jul 20;2022(7):CD014461. doi: 10.1002/14651858.CD014461. PMCID: PMC9298669.
3. Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Chen Q, Maher CG. Red flags to screen for vertebral fracture in people presenting with low back pain. Cochrane Database Syst Rev. 2023 Aug 24;8(8):CD014461. doi: 10.1002/14651858.CD014461.pub2. PMID: 37615643; PMCID: PMC10448864.
4. Han CS, Hancock MJ, Sharma S, Sharma S, Harris IA, Cohen SP, Magnussen J, Maher CG, Traeger AC. Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review. EClinicalMedicine. 2023 Apr 6;59:101960. doi: 10.1016/j.eclinm.2023.101960. PMID: 37096189; PMCID: PMC10121397.
5. Han CS, Maher CG, Steffens D, Diwan A, Magnussen J, Hancock EC, Hancock MJ. Some magnetic resonance imaging findings may predict future low back pain and

disability: a systematic review. J Physiother. 2023 Apr;69(2):79-92. doi: 10.1016/j.jphys.2023.02.007. Epub 2023 Mar 11. PMID: 36914521.

Manuscripts submitted for publication

1. **Han CS**, Hancock MJ, Jones CM, Maher CG. Exploring study level factors which potentially explain differences in results of studies evaluating the STarTBack stratified treatment approach for low back pain: A literature review. Manuscript currently under review in *BMJ Open.*

Presentations (presenter in bold)

1. **Han CS**, Maher CG, Steffens D, Diwan A, Magnussen J, Hancock EC, Hancock MJ. Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review. Oral presentation at the 2022 *Sydney Musculoskeletal Health Annual Scientific Meeting, Sydney*, Australia.
2. **Han CS**, Hancock MJ, Sharma S, Sharma S, Harris IA, Cohen SP, Magnussen J, Maher CG, Traeger AC. Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review. Oral presentation at the *2022 Sydney Musculoskeletal Health Annual Scientific Meeting*, Sydney, Australia.
3. **Han CS**, Hancock MJ, Sharma S, Sharma S, Harris IA, Cohen SP, Magnussen J, Maher CG, Traeger AC. Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review. Oral presentation at the *18th International Forum for Back and Neck Pain Research in Primary Care*, Groningen, The Netherlands
4. **Han CS**, Hancock MJ, Sharma S, Sharma S, Harris IA, Cohen SP, Magnussen J, Maher CG, Traeger AC. Low back pain of disc, sacroiliac joint, or facet joint origin: a

diagnostic accuracy systematic review. Oral presentation at the *2023 Sydney Spine Symposium*, Sydney, Australia (Awarded best oral presentation).

Preface

This thesis is organised into eight chapters and written so that each chapter can be read independently. The University of Sydney allows published papers that arose from the candidature to be included in the thesis.

Chapter One is an introduction to the thesis and provides an overview of the challenges of diagnostic research in low back pain. It also highlights the gaps in the literature in this field that were addressed by the studies in this thesis.

Chapter Two is a perspective paper considering the origins of the term non-specific low back pain, common misconceptions regarding the term, the impact of the term on research and clinical practice, and suggestions for future research to move the field forward. This manuscript is presented as published in *The Spine Journal*.

Chapter Three is a protocol for a Cochrane review to assess the diagnostic accuracy of red flags used to screen for the different types of vertebral fracture in patients presenting with low back pain. This paper is presented as published in the *Cochrane Database of Systematic Reviews*.

Chapter Four is a Cochrane review that investigated the diagnostic accuracy of red flags used to screen for the different types of vertebral fracture in patients presenting with low back pain. This paper is presented as published in the *Cochrane Database of Systematic Reviews*.

Chapter Five is a systematic review to determine the diagnostic accuracy of diagnostic tests to identify the disc, sacroiliac joint, or facet joint as the source of low back pain. This paper is presented as published in *eClinicalMedicine.*

Chapter Six is a systematic review of longitudinal studies assessing whether magnetic resonance imaging can predict future low back pain. This paper is presented as published in *Journal of Physiotherapy.*

Chapter Seven is a literature review that explored study level factors that which may potentially explain differences in study results from trials investigating the StarTBack treatment approach. This paper is presented in the format of the manuscript submitted to *BMJ Open* and is currently under review.

Chapter Eight provides a discussion of the findings of the thesis, including implications for clinicians and recommendations for future research.

Each chapter contains its own reference list; any appendices that were published as online supplementary material are included at the end of the relevant chapter.

Abstract

The overall aim of this thesis is to investigate and address key knowledge gaps in the diagnosis of low back pain and the role of a diagnosis in the treatment of low back pain. **Chapter Two** discusses the origins of the term non-specific low back pain, common misconceptions regarding the term, the impact of the term on research and clinical practice, and suggestions for future research to move the field forward. **Chapter Three and Four** assessed the diagnostic accuracy of red flags used to screen for the different types of vertebral fracture in patients presenting with low back pain. Most red flags were not useful as screening tools to identify those at greater risk of vertebral fracture. However, the four best red flags found were corticosteroid use, older age, trauma, and a contusion or abrasion. **Chapter Five** aimed to determine the diagnostic accuracy of diagnostic tests to identify the disc, sacroiliac joint, or facet joint as the source of low back pain. The results of this systematic review provide evidence that some index tests had informative likelihood ratios for the disc, sacroiliac joint, or the facet joint, challenging the current view that a diagnosis is not possible. **Chapter Six** then assessed whether magnetic resonance imaging can predict future low back pain. The results from this systematic review suggest that some MRI findings have weak associations with future LBP; however, for most of the MRI findings evaluated it was unclear whether there was an association. **Chapter Seven** investigated targeted treatments for low back pain using a classification system (i.e., STarTBack approach) and explored differences in study level factors that may explain differences in results, in studies implementing the STarTBack approach. Reporting of important study level factors in how the STarTBack approach was implemented was often unclear. However, there is some suggestion that key factors may include the individual who implemented the STarTBack tool, whether the recommendations of the tool

were followed, the amount of training the clinician delivering the matched treatment received, and whether clinicians actually delivered the matched treatment.

Chapter One:

Introduction

1.1. Introduction to low back pain

Low back pain is typically defined as pain between the 12th ribs and the buttock creases.[1] Low back pain is a very common condition, and most people will experience low back pain in their life. For example, the one year incidence of experiencing a first ever episode of low back pain is approximately 6.3% to 15.4%, with estimates for the one year incidence of any episode of low back pain being as high as 36%.[2] The one month prevalence of low back pain is estimated to be 23.2% in the general population.[2] The prevalence of low back pain also increases with age[3]. For example, in 2017 the point prevalence estimates of low back pain for 40-44 years old were 12%[4], for 60-64 years old were 20%[4], and peaked at 80-84 years old at 25%[4] and have previously been reported as high as 40%[3]. In Australia, low back pain is ranked third in overall disease burden (disability-adjusted life years), accounting for 3.6% of total disease burden and is the leading cause of burden of all musculoskeletal conditions.[5]

Low back pain results in a large economic burden due to the direct costs (health care) and indirect costs (e.g., social security, and compensation payments). For example, in the United States, low back pain and neck pain combined accounted for USD$134.5 billion in healthcare spending in 2016.[6] In Australia, low back pain cost the Australian healthcare system AUD$4.8 billion annually and was the leading reason for visits to primary care[7], loss of productive years, and exiting the workforce in 2015.[8] There are also large indirect costs contributing to the economic burden of low back pain, such as work absenteeism and loss of productivity. These indirect costs are generally greater than direct costs. A systematic review (n=21 studies) investigated the direct and indirect cost of low back and found that the mean direct costs only accounted for 22% of total costs, indicating that indirect costs are a much larger contributor to the cost of low back pain.[9] In Australia, the total cost of low back pain in 2001 was estimated as AUD$9 billion, however only AUD$1 billion of this was accounted for by direct healthcare

costs.[10] It is widely agreed that low back pain can be very disabling and impose an enormous social and economic burden on the community.[11]

The complexities of diagnosis and treatment of low back pain are likely contributors to the growing burden of disease and costs associated with this condition.[12,13] Performing tests such as magnetic resonance imaging (MRI) for patients with low back pain who have no indications of serious pathology (i.e., for non-specific low back pain) is costly and contributes to overdiagnosis, which increases the risks for patients to receive unnecessary interventions.[14] Studies have shown that liberal imaging is associated with downstream consequences such as subsequent specialist referrals, unnecessary escalation of care, and interventions performed based on imaging results.[13,15] On the other hand, missed diagnoses (e.g., vertebral fracture, cauda equina syndrome) may delay appropriate care and lead to worse patient outcomes and greater costs to the healthcare system. A missed diagnosis may also lead to undertreatment and may contribute to the development of chronic low back pain or more serious disease progression. Chronic low back pain becomes more difficult to treat and requires higher utilisation of healthcare resources, resulting in higher healthcare costs and associated indirect costs secondary to absenteeism from work.[14] A clearer understanding of when pursuing a diagnosis will usefully guide treatment may lead to more informed selection for further testing, with the aim of utilising resources more efficiently to improve patient care.

1.2. Differential diagnosis

Low back pain is a symptom rather than a disease, a symptom which may arise due to several different causes or diseases.[16] It may also be accompanied by pain in one or both legs, with or without associated neurological symptoms.[1,16] For patients presenting with low back pain, most

clinical practice guidelines recommend history taking and physical examination to form a diagnosis[17] and assist in triaging patients into the four broad categories (i.e., serious pathology, radicular syndromes, non-specific low back pain, problem beyond the lumbar spine). The lumbosacral causes of low back pain are often separated into distinct categories, which are described below. For the purposes of this thesis, I will focus on two of the three categories: serious low back pain (e.g., vertebral fracture, infection, and cancer) and non-specific low back pain.

1.2.1. Serious low back pain

Serious causes of low back pain that likely require specific treatments include vertebral fractures, malignancy, spinal infections, axial spondylarthritis and cauda equina syndrome. In the primary care setting, between 1% and 5% of all patients who present with low back pain will have a serious spinal pathology.[16,18-20] The prevalence of serious pathology in patients presenting to the emergency department is thought to be somewhat higher, with one review (n=22; 41320 participants) reporting estimates between 2.5% and 5.1%.[21] However, it is important to note that a large proportion of patients presenting to the emergency department for low back pain are diagnosed with a pathology not originating from the lumbar spine. For example, a recent study (n=1982) using electronic medical record data investigated the proportion of patients presenting to the emergency department with back pain who received a discharge diagnosis of serious spinal or non-spinal pathologies or other musculoskeletal conditions.[22] The study found that 23.9% of patients were diagnosed with a pathology beyond the lumbar spine.[22] A retrospective review of patients presenting to three Australian Emergency Departments between 2016 and 2018 (n=409,405) found that 3.4% (14,024) of patients presented with 'back pain'.[23] Of these 14,024, 54.4% (7631) of patients discharged from the emergency department were diagnosed with a pathology behind the lumbar spine.[23]

In both primary care and emergency department settings vertebral fractures are amongst the most common serious causes of low back pain. According to a systematic review of diagnostic studies (n=14 primary care studies) the median prevalence of vertebral fractures in patients presenting with low back pain was estimated at 3.6% in primary care and 6.5% in secondary and tertiary care.[18] Lumbar vertebral fractures can include different types of fractures such as vertebral traumatic fractures, vertebral stress fractures, and osteoporotic vertebral compression fractures. These three types of vertebral fractures may have different distinguishing clinical histories and clinical features and may require different diagnostic work-up. The three types of fracture have different management, and so a specific diagnosis of the type of vertebral fracture is important.[24]

1.2.2. Radicular syndromes associated with low back pain

Radicular syndromes account for approximately 5-10% of low back pain presentations[25,26] and comprise, radicular pain, radiculopathy and spinal canal stenosis.[16,25] Lumbosacral radicular pain and radiculopathy occur due to spinal nerve or nerve root involvement, however, are distinguished based on clinical presentation. Radicular pain includes dermatomal leg pain, leg pain that is worse than back pain, worsening of leg pain with coughing or sneezing, and positive neurological tension testing (e.g., straight leg raise).[16] Radiculopathy on the other hand is characterised by signs of impaired neurological function such as myotomal weakness, loss of sensation, and/or loss of reflexes. People with low back pain may present with radicular pain and radiculopathy in isolation or in combination.[16]

Spinal stenosis is another lumbar condition and is due to narrowing of the spinal canal. Narrowing of the spinal canal can arise due to a combination of degenerative changes in the

lumbar spine such as facet osteoarthritis, disc bulges, or ligamentum flavum hypertrophy.[16] Characteristic symptoms of spinal stenosis are pain or neurological symptoms exacerbated by walking or extended postures, and relief with flexed postures.[16] It is thought that these symptoms are a result of ischemia of the nerve roots in the cauda equina due to compression.[16]

1.2.3. Non-specific low back pain

Most people presenting to primary care with low back pain of lumbar spine origin will have non-specific low back pain, whereby a specific nociceptive source of pain cannot be identified.[27] The classification of non-specific low back pain implies there is no identifiable pathoanatomical cause for an individual's low back pain (after excluding serious or specific causes).[27] A recent seminar paper in the Lancet highlighted the many potential nociceptive structures in the lumbar spine (e.g., intervertebral discs, facet joints, muscles, ligaments, sacroiliac joint) that may contribute to an individual's pain, despite the challenges in identifying the exact cause for an individual patient.[28]

1.2.4. Potential nociceptive sources of non-specific low back pain

There are many lumbar structures that are plausible sources of low back pain (e.g., disc, facet, sacroiliac joint). For instance, the disc is a commonly cited possible source of low back pain that is susceptible to degenerative, traumatic, or biochemical stressors.[28] Changes occurring in the disc (e.g., disc degeneration, annular fissures) can lead to the growth of nerve endings into the annular fibrosus, and can be sources of nociception.[29,30] Lumbar provocation discography has previously been used for disc stimulation and morphological evaluation and was often used to distinguish a painful disc from other potential sources of pain.[31] A systematic review (n=5 lumbar spine studies) of diagnostic accuracy and utility studies for discography demonstrated that discography is a potentially useful test to diagnose the disc as the source of low back pain.[31]

However, it is important to note the potential harmful effects of the use of lumbar discography (e.g., acceleration of disc degeneration)[32], the potential risk for high false positive rates if control discs are not well controlled, and the limited availability of discography for clinician and patients in primary care must be considered.[31]

The facet joints are a possible cause of low back pain as they are supplied by the medial branches of the dorsal rami which have a large amount of free and encapsulated nerve endings that are nociceptive receptors.[33] Facet joints are susceptible to arthritic changes, degenerative changes, inflammation, and injury; all of which could cause pain to be perceived in the lumbar region.[34] A systematic review (n=17 studies focused on the lumbar spine) highlighted that facet joints can be a pain generator[34] and that subsequent therapeutic application of controlled local anaesthetic blocks or medial branch blocks to the facet joint can relieve these symptoms. Therefore, the application of these blocks can be used to diagnose facet joint pain.[35] For example, a systematic review of diagnostic studies (n=17 studies) that utilised dual diagnostic blocks with at least 75% pain relief demonstrated the mean prevalence of facet joint pain ranged between 16 to 41%.[35]

The sacroiliac joint is another potential source of low back pain which is highly innervated including nociceptive receptors.[29] The sacroiliac joint is considered to have intra-articular or extra-articular causes of pain. For example, the intra-articular causes could be infection or inflammation, and the extra-articular causes could be enthesopathy, ligament injury, or myogenic injury.[29] These innervated structures of the sacroiliac joint have been shown to be relieved with local anaesthetic, and therefore the International Association for the Study of Pain (IASP) recommends the use of single or double blocks for diagnosing sacroiliac joint pain.[29]

With the challenges of identifying a specific nociceptive cause of low back pain, there has recently been greater focus on factors contributing to an individual's pain and associated disability (e.g., psychological, genetic, biophysiological, social, and lifestyle factors)[16] and less focus on pathoanatomical factors. Low back pain is associated with a range of biophysical, psychological, and social factors that influence one's function, societal participation, and personal financial prosperity.[16] It is important to acknowledge the complexity of low back pain and that an individual's pain and associated disability is likely a combination of several factors mentioned above.

1.3. Diagnosis of serious low back pain

Although it is uncommon for presentations of low back pain to be due to a serious underlying condition (Maher 2017), it is important that health professionals can identify or screen for serious pathology using red flags. The accuracy of red flags is important. On the one hand, if the tests have a high rate of false alarms, people without a spinal fracture may undergo unnecessary imaging. On the other hand, missing a spinal fracture will result in a delay in receiving treatment and may lead to worse outcomes. Therefore, robust data on the diagnostic accuracy of red flags to screen for vertebral fractures is needed.

1.3.1. Reference standards (i.e., imaging tests) for serious low back pain

Imaging procedures are sometimes used as the reference standard to identify serious pathologies such as vertebral fractures.[20] Imaging techniques such as x-ray, MRI, CT, and SPECT are often used when there is an increase in suspicion of vertebral fracture when red flags are present (e.g., positive index tests such as trauma).

Most clinical practice guidelines recommend history and physical examination to identify low back pain patients with a higher likelihood of vertebral fracture.[17] Those who require further diagnostic work up may require different imaging procedures for the different types of vertebral fracture. However, imaging can be costly and has been shown to not necessarily lead to better outcomes.[13] Therefore, informative index tests are needed to avoid unnecessary imaging.

The American College of Radiography (ACR) recommends standard radiography for initial imaging in patients presenting with low back pain with a suspicion of osteoporotic vertebral compression fracture.[36] Genant's semi-quantitative assessment and the algorithm-based qualitative approach are methods used to classify and confirm the presence of a vertebral fracture on standard radiographs.[37,38] Computed tomography (CT) is recommended over a plain radiograph if there is suspicion of vertebral traumatic fracture or vertebral stress fracture.[36]

Magnetic resonance imaging (MRI) is now preferred over CT in evaluating patients with suspicion of lumbar stress fracture due to radiation exposure concerns.[39] MRI can detect bone marrow oedema early, facilitating early diagnosis of lumbar stress reaction and fracture without radiation exposure.[39,40] Given that most patients presenting with low back pain with suspicion of lumbar stress fracture tend to be from younger age groups, MRI may be beneficial as first-line imaging given its high sensitivity and specificity, and no ionising radiation exposure.[40-43]

1.3.2. Index tests (i.e., clinical examination tests) for serious low back pain

Most clinical guidelines recommend the use of red flags (e.g. unexplained weight loss) to assist clinicians to identify cases of serious pathology[44] However, guidelines are inconsistent regarding which red flags to use, and many endorsed red flags that are of limited or unknown diagnostic accuracy.[27] An example is the red flag ‘thoracic pain’ which has a positive and

negative likelihood ratio of 1.01 for detecting malignancy in patients attending primary care for management of low back pain.[18,45]

The lack of diagnostic studies is well illustrated by the 2013 Cochrane review of red flags for spinal malignancy in patients presenting with low back pain[46] which found only eight studies. The review (n=8 studies; 7361 participants) found the diagnostic accuracy of most red flags was poor whether in isolation or combination, except for 'history of cancer' which demonstrated informative +LR ranging from 12.7 to 35.0.[46] For cauda equina syndrome, a recent systematic review (n=7; 569 participants) which only included seven studies found that only 'saddle anaesthesia' (+LR 2.00) demonstrated the most informative pooled diagnostic value.[47] For vertebral fracture, a previous Cochrane review[20] (n=8 studies; 7378 participants) found three red flags were informative: such as significant trauma (+LR range: 3.42 to 12.85), older age (+LR range: 3.69 to 9.39), and corticosteroid use (+LR range: 3.97 to 48.50).[20] However there is uncertainty on the diagnostic accuracy of red flags for vertebral fracture as there is a limited number of diagnostic accuracy studies, results were based on single studies, and there was no differentiation between different index tests for the different types of vertebral fracture.[20] There is a paucity of primary diagnostic research in this field and therefore red flags for vertebral fracture (and different types of fracture) are further investigated in Chapter Four of this thesis in a Cochrane review.

1.4. Diagnosis of non-specific low back pain

There is a common view that a pathological diagnosis is not possible, and that diagnostic research in this field has been completed thoroughly.[48] However, it is important to recognise that little research has been conducted in this field. For example, the Hancock et al 2007 systematic review of diagnostic studies to identify the source of low back pain only identified

41 studies in total.[49] Most of the included studies in this review[49] were of low quality, had a small sample size, and the primary focus of some of the included studies was not the evaluation of diagnostic accuracy.[49] In contrast, a recent Cochrane review investigating exercise therapy for non-specific low back pain included 249 trials[50] and a systematic review investigating psychological interventions for non-specific low back pain included 97 trials.[51] It is clear there is a need for more studies focusing on the available diagnostic tests for low back pain compared to studies investigating treatments for non-specific pain.

Diagnostic tests such as imaging need to be able to increase (when present) or decrease (when absent) the likelihood of a particular structure (e.g., disc) being the source of pain.[49] A diagnosis based upon these tests will only have utility if the diagnosis is able to inform one or more of the following: prediction of prognosis, understanding disease progression, or identifying those who respond to specific treatments. For the purposes of this thesis, the index tests i.e., clinical examination and imaging tests have been presented separately below.

1.4.1 Reference standards (i.e., discography and diagnostic injections) for non-specific low back pain

There is some controversy surrounding the gold standard for the diagnosis of the disc, facet joint, or sacroiliac joint (SIJ) as the source of LBP; however, there are some tests considered as acceptable reference standards. A reference standard is a procedure that is considered the best available method of classifying a person as having or not having a health condition.[52] There is a general consensus on which reference standards to use for the disc, facet joint, and SIJ as recommended by the International Association for the Study of Pain[53] and the Dutch Multidisciplinary Guidelines.[29] For example, for the disc, discography is recommended as a reference standard for discogenic low back pain.[29,53] For the facet joint and SIJ, single or double

anaesthetic injections are recommended.[53] Most reference standards used to identify the source of low back pain are not suitable nor available for routine use in primary care, therefore identifying simpler index tests to help identify the source of low back pain is necessary.

The accuracy of the reference standards will impact the estimates of diagnostic accuracy of index tests compared to these reference standards. Some studies that have used discography as a reference standard to identify the disc as the source of pain have found a high rate of false positive responses with discography when comparing symptomatic and asymptomatic discs.[54,55] Studies have also found low false positive rates when a stricter criterion for a positive response to discography is used (i.e., reproducible pain greater than or equal to 6/10, 50 psi intradiscal pressure, 3.5 mL total volume).[54,55] Unfortunately, not all studies follow the same criteria for a positive response. There is limited primary diagnostic data for a sensitivity analysis to investigate whether diagnostic accuracy values for index tests differ depending on whether a stricter or looser criterion for a positive response is used.[49]

1.4.2. Index tests (i.e., imaging tests) for non-specific low back pain

A diagnosis based on imaging findings such as disc degeneration in patients with low back pain is controversial. The controversy largely arises because changes on imaging can be present in both symptomatic and asymptomatic individuals[56] and therefore the clinical relevance has been challenged. However, previous reviews have been limited by very few primary studies with enough diagnostic data to draw definitive conclusions.[49]

There is currently limited diagnostic research investigating imaging index tests to identify the source of low back pain, however, there has been some interesting research in this field. A systematic review by Hancock et al 2007 (n=41 studies) investigated the disc, facet, or

sacroiliac joint as the source of pain.[49] This review found that some imaging tests for the disc such as MRI evidence of disc degeneration, high intensity zone, and Modic changes, increased the likelihood of the disc being the source of low back pain. For the sacroiliac joint, radionuclide imaging increased the likelihood of the sacroiliac joint being the source of low back pain. No imaging tests for the facet joint were found to be informative. Contrary to previous literature, a diagnosis may be possible for patients with persistent low back pain[49], however further research is needed. Therefore, given the Hancock et al 2007 review[49] is now over 15 years old, this is investigated in Chapter Five in a systematic review.

1.4.3. Index tests (i.e., clinical examination) for non-specific low back pain

A highly cited diagnostic systematic review by Hancock et al 2007 (n=538 cites) suggested that there was a lack of clinical tests that were able to identify a particular structure or nociceptive cause of low back pain (e.g., the disc, sacroiliac joint, or facet joint).[49] However, this systematic review[49] was limited by low study quality of included studies, small sample sizes, and the primary focus of some of the studies was not for diagnostic accuracy purposes.[49] This systematic review is also now outdated and therefore is updated in Chapter Five.

The Hancock et al 2007 review located 41 studies that Investigated tests (e.g., physical examination and imaging findings) to identify the disc, facet joint, and sacroiliac joint (SIJ) as the source of low back pain.[49] Of these 41 studies, 27 studies investigated the disc as the source of low back pain, 7 studies investigated the facet joint, 6 studies investigated the sacroiliac joint, and 1 study investigated all three sources.[49]

The Hancock et al 2007 review found that there were some informative diagnostic tests to increase or decrease the likelihood of a specific diagnosis for low back pain.[49] For studies

investigating the disc, the only clinical tests that were found to increase the likelihood of the disc being the source of pain was centralisation phenomenon (+LRs 2.80, 95%CI: 1.40 to 5.30) (i.e., repeated end-range movements that result in distal pain originating from the spine progressively moving towards a central location)[57]. For the sacroiliac joint, a combination of pain provocation tests increased the likelihood of the sacroiliac joint being the source of pain (+LR 3.20; 95%CI: 2.30 to 4.40). No clinical examination tests for the facet joint were found to be informative. Another systematic review (n=5) which included only five studies found a combination of sacroiliac joint provocation tests increased the likelihood (+LR 2.13; 95%CI: 1.20 to 3.90) of the sacroiliac joint being the source of pain.[58] In the last 15 years, only two systematic reviews[49,58] have explored clinical tests to identify the source of low back pain. Systematic reviews are limited by a small number of primary studies specifically investigated physical examination findings to identify the source of low back pain, and instead many primary studies have focused on imaging findings.

1.5. Consequences of a diagnosis

Ultimately, a diagnosis will only have utility if the diagnosis is able to predict an individual's prognosis or response to a particular treatment. Inappropriate or excessive use of tests such as imaging in search of a diagnosis can result in negative consequences. There are, however, consequences to not using the available tests that have some scientific rigour in clinical practice. This has the potential for missed diagnosis of serious pathology (e.g., vertebral fracture) or failing to identify patients at risk of persisting non-specific low back pain. This will be explored further below.

1.5.1. Consequences of diagnosis for serious low back pain

The use of clinical tests to identify vertebral fracture aims to minimise both overtreatment and undertreatment. Identifying the different types of vertebral fracture as a cause of low back pain is crucial as the management of these fracture types is different, and markedly different to cases of non-specific low back pain.[20] Without appropriate index tests to increase or decrease the suspicion of a vertebral fracture, patients may undergo unnecessary imaging or risk missed diagnosis.[13] Some of these imaging modalities lead to excessive costs, potentially unnecessary radiation exposure, unnecessary patient anxiety, and can also lead to negative downstream consequences (e.g., more invasive procedures).[13,59] Overdiagnosis and unnecessary imaging also leads to the issues of more invasive and unnecessary treatments, with previous research demonstrating that more care does not lead to better outcomes and may even lead to poorer outcomes.[13,60]

On the other hand, missing a spinal fracture will result in a delay in receiving appropriate treatment and reduce quality of life. The potential for misdiagnosis of vertebral fracture without appropriate clinical tests is further complicated by the low prevalence of vertebral fracture in primary care. However, without robust evidence for the utility of index tests, clinicians will continue to face the challenge of being selective when referring patients for further investigations to rule out vertebral fracture.

1.5.2. Consequences of diagnosis for non-specific low back pain

Providing an accurate pathoanatomical diagnosis in individuals currently categorised as having non-specific low back pain is challenging.[49] It is possible that the inability to identify a pathoanatomical cause of low back pain is a contributing factor to unnecessary imaging, over treatment and generally small effect sizes for most interventions.[27,60] The label of non-specific

low back pain may contribute to non-specific treatments, largely targeting the symptoms. For example, many current clinical practice guidelines recommend broad treatment approaches that have been found to demonstrate a small beneficial effect such as exercise, medication, or manual therapy for patients with non-specific low back pain.[61] However, guidelines provide very little to no information on treatment selection and how to individualise these treatments.

The ability to form a diagnosis in people with low back pain is an important step towards potentially moving beyond the non-specific label and developing new, more targeted, and specific treatment approaches. For example, the pathology and causal mechanism(s) driving Modic type 1 changes has been linked to bacterial infection, inflammation, and bone marrow oedema.[62-64] Treatments have been proposed to address these causes and mechanisms (e.g., antibiotics, Zoledronic acid or Denosumab for Modic type 1 changes), however we require much further research evaluating their safety and effectiveness. Specific treatment approaches targeting a pathology or causal mechanism offer research opportunities to potentially develop targeted and more effective treatments in the future.

1.5.3. Consequences of diagnosis for prognosis of non-specific low back pain

A diagnosis may also provide insights into prognosis and natural history of disease progression. The limited evidence in this field suggests there may be some association between MRI findings and future low back pain, providing some insight into the prognosis when certain pathoanatomical structures are present. For example, a systematic review by Steffens et al 2014 included longitudinal studies (n=12 studies) to investigate the relationship between MRI findings and future low back pain.[65] Definitive conclusions were not possible in this review due to limitations such as a paucity of high-quality studies and significant heterogeneity between studies.[65] However, single studies did report significant associations between the

presence of Modic type 1 with future pain, and disc degeneration with future disability, in those with low back pain.[65] Identifying MRI findings and their relevance to low back pain could assist clinicians to define subgroups that follow a similar natural course or that moderate the effects of various treatments. Given the limitations of the Steffens et al 2014 review, the relationship between MRI findings and future pain and disability is addressed in Chapter Six in a systematic review.

1.6. Management of low back pain

The management of low back pain will be dependent on whether a patient's pain is arising from a serious cause or from non-serious causes, usually labelled as non-specific low back pain. For the purposes of this thesis, I will only present on the management of vertebral fractures (and not the other serious causes of low back pain) as only vertebral fracture is relevant to this thesis. The different treatment approaches currently available are described below.

1.6.1. Current management of vertebral fracture

Fractures of the lumbar spine are conventionally managed conservatively, however in some complex cases may require surgery (e.g., high trauma).[66] Clinical practice guidelines recommend that the management for acute vertebral fractures involves pain management, short term bed rest, early graded exercise (e.g. physiotherapy), and spinal braces (spinal orthosis).[66-68] A prospective study (n=259) showed that patients with vertebral fracture who demonstrated improvements in pain and disability within the first 3 weeks of conservative management had a 95% chance of maintaining these improvements for up to 12 months.[69] The Clinical Practice Guidelines for vertebral fracture in 2017 demonstrated that there are few high quality studies available and many of the recommendations were based on limited evidence and therefore

inconsistent in their recommendations.[68] There is currently no definitive consensus on the best conservative approach, however it is widely agreed that conservative management should be first line treatment before any surgical intervention.[66]

Percutaneous vertebral augmentation (e.g., vertebroplasty or kyphoplasty) is often considered in patients with persistent pain affecting their quality of life and function, despite conservative management.[69-73] However, the recommendations for vertebral augmentation were based on weak evidence, and high quality evidence from the American Academy of Orthopaedic Surgeons now recommends against vertebroplasty for vertebral compression fracture.[74-77] Kyphoplasty is still recommended as an option for those with osteoporotic vertebral fracture with no associated neurological deficits as per the American Academy of Orthopaedic Surgeons.[78] There remains uncertainty in the current recommendations for conservative or interventional management of vertebral fracture.[66,68]

1.6.2. Current management of non-specific low back pain

Treatments for non-specific low back pain largely aim to reduce pain and its consequent disability primarily through generic (i.e., non-targeted) pharmacological and/or non-pharmacological treatments. Clinical practice guidelines vary a little between countries, however most clinical guidelines recommend similar approaches to the management of low back pain.[27] The main components of guideline recommendations consist of education, reassurance, analgesic medication, and non-pharmacological treatments (e.g., manual therapy, exercise).[27,61] However, reviews have shown some of these pharmacological and non-pharmacological treatments generally only have small effects.[27,61]

1.6.3. Management of acute non-specific low back pain

For acute low back pain, clinical practice guidelines usually recommend the prescription of simple analgesics before prescribing more powerful analgesics.[79] However, an issue is that many individuals with low back pain continue to be prescribed common medications that have been shown to be ineffective and carry potential risks for adverse reactions.[80] A clear example is the use of paracetamol for low back pain. Paracetamol has been shown to be ineffective for acute low back pain and is yet to be rigorously tested for chronic low back pain. A recent overview of systematic reviews (n=36 reviews) found strong evidence that paracetamol is ineffective in reducing acute low back pain (mean difference 0-10 point scale: 0.2; 95%CI: –0.1 to 0.4 and is also linked with adverse effects (e.g., elevated liver enzymes).[80]

More powerful analgesics such as opioids are also commonly prescribed for low back pain, however, their efficacy and safety have also been challenged. For example, a systematic review (n=42 trials; 6128 participants) found that stronger analgesics such as opioids only demonstrated a small improvement in pain when compared to paracetamol (mean difference 0-100 scale: -6.7; 95%CI: -11.9 to -1.5) and placebo (mean difference 0-100 scale: -6.3; 95%CI: -10.5 to -2.2) in the short term (2-hour follow-up).[81] This systematic review also found opioids were no more effective than simpler analgesics such as NSAIDs, however opioids were associated with greater rates of adverse events compared to paracetamol, placebo, and NSAIDs.[81] A recent Cochrane review (n=32) found that NSAIDs were slightly more effective than placebo in the short term for pain and disability reduction, however effects were unlikely to be clinically important.[82] The widespread use of pharmacological treatments needs to be reconsidered given its limited efficacy and safety, and other forms of treatment should be considered.

Non-pharmacological treatments for acute low back pain include treatments such as manual therapy, acupuncture, and advice to stay active.[27,61] However, many guidelines do not inform researchers and clinicians on how to individualise non-pharmacological treatments and/or how to identify those who will likely benefit the most.[27] One approach is to only consider these treatments if first line care (i.e., education, reassurance, simple analgesics) is ineffective as approximately 50% of patients with acute low back pain will recover within 3 weeks with minimal intervention.[83] Although non-pharmacological treatments are not associated with many adverse reactions, not all patients with acute low back pain will likely need or benefit from these treatments. These treatments may be an unnecessary use of healthcare recourses given the good natural history of acute low back pain.[27]

1.6.4. Management of chronic non-specific low back pain

For chronic low back pain, the clinical practice guidelines largely recommend non-pharmacological treatments such as education, reassurance, exercise, and an increased awareness of the influence of psychosocial factors on a patient's recovery.[79] Exercise is one of the most common treatment approaches for chronic low back pain and has been shown to be one of the most effective treatments. A recent Cochrane review (n=249 trials; 24486 participants) found exercise when compared to placebo demonstrated a pooled effect of 11.8 points (95%CI: 1.6 to 22.0) on a 0-100 pain scale and a pooled effect of 5.2 points (95%CI: 0.8 to 11.3) when compared to other conservative treatments at long term (>12 month) follow-up.[50] The effect sizes were smaller for the outcome of disability.[50] However, chronic low back pain is accepted as a complex multifactorial biopsychosocial condition[16], and often treatments such as exercises should be incorporated as part of an individualised approach and not necessarily a standalone treatment.

Clinical practice guidelines recommend clinicians take into consideration both the biophysiological and psychosocial contributors in the management of chronic low back pain[84]. However, most current treatment approaches are unimodal and fail to address the various and complex factors contributing to an individual's low back pain. This may be why the effects of most recommended treatments for chronic low back pain are modest and short lived.[85] Combined approaches (e.g., exercise and psychological therapies) may potentially elicit greater treatment effects if implemented with a patient-centred approach addressing the factors contributing to an individual patient's pain, and consequent disability.[51,85,86]

Cognitive functional therapy (CFT) is a patient-centred approach that aims to address the complex nature of chronic low back pain through facilitating self-management, by targeting the individual's pain-related behaviours and emotions contributing to their pain.[85] There was been some initial promising results from two small trials comparing CFT to manual therapy and exercises (n=121)[87], and CFT to pain education and group-based exercises (n=206)[88]; however both trials had high rates of loss to follow-up. More robust evidence comes from a However, a recent three-arm RCT (n=492)[85] assessed the effectiveness of individualised CFT, delivered with or without movement sensor biofeedback, compared to usual care for patients with chronic low back pain. This three-arm trial found that CFT alone (mean difference: 4.6; 95%CI: 3.4 to 5.9 on a 24-point scale) and CFT plus biofeedback (mean difference: 4.6; 95%CI: 3.3 to 5.8 on a 24-point scale) both resulted in clinically important effects for disability at 13 weeks follow-up when compared to usual care.[85] Importantly the effects were sustained at 12 months (mean difference: 4.8; 95%CI: 3.5 to 6.0 and 5·4; 95%CI: 4.1 to 6.6 on a 24-point scale, respectively).[85] Targeted, individualised treatment approaches such as CFT highlight the importance of developing and assessing the efficacy of targeted treatments to push the field forward.

Pharmacological approaches for chronic low back pain are currently not well supported in the literature despite their continued use and should be used with caution. For example, the 2016 NICE draft guideline recommended the use of oral NSAIDs only and advises against opioids and paracetamol for chronic low back pain.[84] Simpler analgesics such as paracetamol have not been rigorously assessed for chronic low back pain.[80] Stronger analgesics such as opioids may provide short term pain relief based on very low certainty evidence, however the potential harms of opioids suggest caution in their use.[81,89] A systematic review (n=16 RCTs) comparing opioids to placebo demonstrated a small statistically significant improvement (standardised mean difference: -0.41; 95%CI: -0.48 to -0.34) in the intermediate term (12-15 weeks follow-up), however in the long term (>6 months) adverse events were much more prevalent in the opioid group and treatment effects were not statistically significant (risk ratio: 0.84; 95%CI: 0.63 to 1.12).[89] The 2016 NICE draft guideline also recommends against the use of anti-depressants or neuromodulators such as gabapentin or pregabalin.[84] A recent overview of systematic reviews (n=26 reviews; >25,000 participants) found that anti-depressants were not effective (<10-point change on 100-point scale pain) for low back pain amongst many other health conditions.[90] For anti-convulsant medications, a systematic review (n=9; 859 participants) found that these medications were ineffective for the treatment of low back pain, and also found high quality evidence of higher adverse health risks when using these medications.[91] Pharmacological treatments for chronic low back pain should be considered very cautiously given their lack of efficacy and risk for adverse events.

If primary care treatments such as exercise and psychological therapy are ineffective for an individual with low back pain, other management approaches such as multi-disciplinary care could be considered. Multi-disciplinary care involves input from a variety of clinicians such as

physiotherapists, psychologists, pain specialists, with consideration of lifestyle factors.[16] Multidisciplinary approaches are endorsed in guidelines by American College of Physicians and American Pain Society[92] and the 2016 NICE draft guideline[84]. A Cochrane review (n=41; 6858 participants) found that a multi-disciplinary treatment approach was only slightly more effective for reducing pain, disability, and improving work status when compared to usual care.[86] However, multi-disciplinary care is not cost-effective when compared to other treatments (e.g., usual care) and is often associated with significantly higher costs than other forms of treatment.[93] Therefore, multi-disciplinary care should not be considered for first line care for chronic low back pain given the current evidence.

1.6.5. Stratified care for low back pain

Given the challenges and costs associated with managing chronic low back pain, and the difficulty of identifying a pathoanatomic diagnosis, the ability to identify those at different risk levels of experiencing poor outcomes may be important, in attempts to match patients to the right care (and amount of care) based on their risk level. Tools such as the Keele STarT Back Screening Tool aim to identify patients with an increased risk of delayed recovery, and to intervene early with more intensive therapy rather than wait for first-line treatment to fail.[94,95] Initial studies from Hill et al 2011[95] and Foster et al 2014[94] demonstrated small positive results though only the Hill et al 2011[95] study demonstrated statistically significant results (change in NRS 0-100 scale: 5.5; 95%CI: 2.3 to 8.6 and 2.9; 95%CI: 2.3 to 3.5 points, respectively). The results of both studies[94,95] suggest that the use of the STarTBack Tool may improve pain and disability, reduce time off work[94], and improve healthcare efficiency[95]. However, follow-up studies have not been able to replicate the earlier results, with many of the follow-up studies demonstrating no difference in outcomes.[96-104] Given the negative results of the replication studies, the stratified approach to low back pain has recently been challenged. A recent review

(n=58 studies) found insufficient evidence to support the use of stratification systems (e.g., STarTBack Tool) compared to non-targeted interventions for non-specific low back pain.[105]

There are several possible reasons for the variable findings across studies investigating the STarTBack approach. This may potentially be explained by aspects of the design and/or conduct of those subsequent studies.[96-104] For instance, the STarTBack Tool and/or matched treatment approach may not have been implemented in the subsequent studies in a similar manner to the original studies.[94,95] Chapter Seven is a literature review exploring the differences in study level factors that may explain differences in study results, in studies that have tested the STarTBack approach for the management of low back pain.

1.6.6. Targeted treatments for non-specific low back pain

Targeted treatments for low back pain are a controversial topic and very little high-quality research has been done on this topic. One potential way to provide targeted treatments might be to target a presumed nociceptive or pathoanatomical source (e.g., Modic changes) of low back pain, with the aim to improve patient outcomes and provide a treatment with long lasting effects. However, a key challenge is whether treatments targeting a specific nociceptive or pathoanatomical source will lead to better patient outcomes.[106] The current available evidence on treatments targeting the various potential nociceptive causes of low back pain is explored below.

A potential source of low back pain is the disc and some studies have attempted to identify the pathoanatomical changes in the disc with the aim to specifically target these changes during treatment. For example, Modic type 1 changes are a potential nociceptive source of low back pain, and some efforts have been made to identify the pathology (e.g., bacterial infection,

inflammation, and bone marrow oedema) driving the nociceptive cause.[62,64] A controversial RCT (n=162) investigated the effectiveness of antibiotic treatment for patients with low back pain and Modic type 1 changes. The authors reported that patients treated with antibiotics compared to placebo at 1 year follow-up demonstrated improved pain (pain intensity 0-10: 3.7; 95%CI: 1.3 to 5.8 vs 6.3; 95%CI: 4.0 to 7.7 points respectively) and disability (RMDQ: 7.0; 95%CI: 4.0 to 11.0 vs 14.0 95%CI: 8.0 to 18.0 points respectively).[62,63] These treatment effects would be considered large, however a replication study was not able to reproduce these results.[107] The replication trial (n=180) comparing antibiotics to placebo for Modic changes (type 1 or 2) demonstrated no difference between groups in pain (pain intensity 0-10: 0.3; 95%CI: 0.3 to 0.8) or disability (RMDQ: 1.9; 95%CI: 0.7 to 3.2) outcomes at 3 month follow-up.[107] A follow-up analysis investigating the effects of antibiotics compared to placebo for those with Modic type 1 changes only found slightly larger effects at one year follow-up (RMDQ: 2.3; 95%CI: 0.4 to 4.2).[107] A second replication trial from Iraq has now been withdrawn.[108]

Zoledronic acid has been explored as another targeted treatment option for patients with Modic changes. A proof of principle study (n=103) initially showed that a one-off treatment of zoledronic acid demonstrated no statistically significant improvement in LBP compared to placebo (VAS 0-100: 8.2; 95%CI: -2.4 to 18.8) for patients with Modic changes.[64] However, this study's[64] biggest limitation was blinding to the treatment received. A recent pilot RCT (n=25) demonstrated that zoledronic acid compared to placebo at 4-week follow-up had small improvements on pain (NRS intensity 0-10: 5.1 ± 1.9 vs. 6.9 ± 1.8 points, respectively).[109] It is important to note that the sample size in the replication study was small (n= 25) , improvements in pain between groups were not seen at longer term follow-up, and in the original study pain

improvements were not seen on the visual analogue scale (equivalent to the numeric rating scale; NRS).[109]

Another controversial treatment for low back pain is intradiscal methylene blue injection. Methylene blue was thought to have a neuro-ablative or possibly anti-inflammatory effect on the disc through its action on nitric oxide.[110] Based upon this rationale, a RCT (n=72) investigated the effectiveness of methylene blue injection for low back pain targeting the annular fissure.[111,112] The results of this RCT demonstrated methylene blue injection was significantly more effective compared to sham treatment (NRS 0-100 scale: 38.6; 95%CI: 31.5 to 45.6) at 6 months follow-up.[111] However, a high-quality replication RCT (n=84)[113] was unable to duplicate these results and found no statistically or clinically significant differences between groups (NRS 0-100 scale: 2.0; 95%CI: -8.0 to 12.0).

The sacroiliac joint is another potential source of low back pain. Sacroiliac joint intra-articular and extra-articular steroid injections have been used as a form of targeted treatment for those with suspected sacroiliac related low back pain. A systematic review (n=11 studies; 6 RCTs, and 5 non-RCTs) investigated the effectiveness of therapeutic interventions for the sacroiliac joint in patients with low back pain.[114] Of those 11 studies, four studies evaluated intra-articular injections of which three studies were non-randomised studies. Statistical pooling was not possible in this study due to the heterogeneity in the data of the included studies in this review[114], and due to the limited and poor-quality evidence it was not possible for the review to recommend the use of sacroiliac joint injections.[114]

Considering the lack of high-quality diagnostic research and limited evidence there appears to be substantial overuse of treatments targeting specific lumbar structures, such as facet and

sacroiliac joint injections. For example, a study assessing the utilisation of interventional management techniques in the US Medicare population reported an increase from 425,000 interventions in 2000 to 2.2 million interventions in 2013, demonstrated a 417% increase in facet and sacroiliac joint injection uptake.[115] Fluoroscopic guidance has been recommended when performing sacroiliac injections, however there is no consensus on what dose or type of steroid will provide the best therapeutic effect.[116,117]

Radiofrequency ablation is another treatment targeted at the sacroiliac joint. A systematic review (n=11 RCTs; 253 participants) investigated radiofrequency ablation for chronic low back pain based on different causes of low back pain (e.g., sacroiliac joint).[118] In this review, only two RCTs (n=48 participants) investigating radiofrequency ablation for sacroiliac joint pain were included.[118] Of these two RCTs, results for radiofrequency ablation were only reported in one study, which reduced pain (NRS 0-10 scale: 2.4; SD 2.3) when compared to sham treatment.[118] The review was limited to two RCTs with a small number of participants (n=28 and 51 respectively).[118] The MINT trials (n=228 participants in sacroiliac joint trial) demonstrated that an exercise program with the addition of radiofrequency denervation showed a small statistically significant improvement in pain (NRS 0-10 scale: 0.71; 95%CI: 0.06 to 1.35) at 3 month follow-up for those with low back pain thought to be originate from the sacroiliac joint.[119]

The facet joint is another potential source of low back pain that could be addressed with. radiofrequency denervation. A Cochrane review[120] (n=23 RCTs; 1309 participants) investigated radiofrequency denervation for chronic low back pain and included 12 RCTs assessing radiofrequency denervation for facet joint pain. Overall, moderate quality evidence suggests that facet joint radiofrequency ablation has a greater effect on pain compared with

placebo over the short term (mean difference 0–10-point scale: -1.46; 95%CI: -2.28 to -0.67).[120] The Cochrane review demonstrated the lack of high-quality evidence and lack of primary data.[120] However, the 2016 NICE draft guidelines endorsed the use of this treatment modality.[84] The MINT trials (n=251 participants in facet joint trial) demonstrated that an exercise program with the addition of radiofrequency denervation did not improve outcomes, nor was it cost effective at any follow-up time point for those with low back pain thought to be originating from the facet joint.[119]

Treatments that aim to target the specific nociceptive causes of low back pain (e.g., antibiotics, Zoledronic acid, Denosumab, radiofrequency ablation) require more and higher-quality research regarding their safety and effectiveness. The general consensus in the literature is that non-specific low back pain has no identifiable pathoanatomical cause, and for this reason, there is very limited research investigated specific treatment approaches or treatment responses based on certain pathoanatomical structures. However, further research to better understand the pathoanatomical structures and causes and mechanisms driving nociception are needed before we assess and test interventions targeting specific structures or nociceptive sources. Assessing and implementing interventions that target a specific nociceptive source or structure prior to gaining knowledge and understanding of the causes and mechanisms behind the pathoanatomical structures, is unlikely to clarify nor move this field of research forward. These research approaches present an opportunity to develop and assess treatments that target a specific cause of non-specific low back pain.

1.6.7. Possible future interventions with greater and longer lasting effects

In other areas of medicine, advances in effective interventions often follow from a better understanding of the causes and mechanisms of a condition or disease, however this has proven

challenging in the low back pain field.[121] Ultimately, the aim of a deeper understanding of the nociceptive or pathoanatomical causes and mechanisms behind a condition or disease is to develop interventions targeting such causes. The objective of having a deeper understanding of the nociceptive or pathoanatomical causes and mechanisms is to then develop interventions with longer lasting effects and improve long term patient outcomes. There are several approaches that may be helpful to move the field beyond non-specific interventions for low back pain to improve patient outcomes.

A greater understanding of the pathoanatomical cause(s) of low back pain may lead to new management approaches that are able to directly target the specific pathoanatomical and nociceptive sources contributing to low back pain. For example, rheumatoid arthritis was once a debilitating condition leading to deformities and joint damage, however, advancements in understanding pathological mechanisms behind rheumatoid arthritis have led to the development of more targeted, more effective, and safer treatments radically changing the long term outcomes with fewer adverse reactions.[122-124]

Future research should place greater attention on investigating the pathoanatomical causes of low back pain and continue research where some promising findings have emerged. For example, Modic type 1 changes have been shown to be more prevalent in people with low back pain[56,125] and there is also some evidence suggesting those with Modic type 1 changes have worse prognosis.[126] However, a better understanding of the aetiology and pathogenesis of Modic type 1 changes (e.g., infection, inflammation, trauma) may allow researchers to explore more specific and targeted treatments to improve long term patient outcomes. Advancements in medical technology and imaging in the future may present new research opportunities to explore the pathoanatomical and nociceptive causes and mechanisms further and help push the

field forward, to move beyond non-specific treatments. However, to utilise these advancements, we must first have a better understanding of the role that diagnostics in low back pain has in research, and in clinical practice.

1.7. Aims of thesis

The broad aim of this thesis was to identify and address key knowledge gaps in the diagnosis of low back pain and the role of a diagnosis in the treatment of low back pain. The knowledge gaps identified included the misconceptions of the term non-specific low back pain and the available diagnostic tests to screen for vertebral fracture and identify the causes of non-serious low back pain. I also identified the lack of research surrounding the prognostic relationship between imaging tests and low back pain and specific treatments available for low back pain using classification systems. A variety of different methods were used to address this aim, including systematic reviews with meta-analysis, Cochrane review, literature review, and a perspective paper.

The specific aims of this thesis were:

1. Explore the origins of the term non-specific low back pain, the misconceptions of the term, the impact of the term on research and clinical practice, and provide suggestions for future research to move the field forward (Chapter Two);

2. Investigate the diagnostic accuracy of red flags (index tests) used to screen for the different types of vertebral fracture in patients presenting with low back pain (Chapter Three and Four);

3. Investigate determine the diagnostic accuracy of diagnostic index tests to identify the disc, sacroiliac joint, or facet joint as the source of low back pain (Chapter Five);

5. Investigate whether magnetic resonance imaging can predict future low back pain (Chapter Six);

6. A Explore the study level factors that have tested the STarTBack treatment approach for the management of musculoskeletal (Chapter Seven).

1.8. References

1. Dionne CE, Dunn KM, Croft PR, et al. A consensus approach toward the standardization of back pain definitions for use in prevalence studies. *Spine (Phila Pa 1976).* 2008;33(1):95-103.
2. Hoy D, Brooks P, Blyth F, Buchbinder R. The Epidemiology of low back pain. *Best Pract Res Clin Rheumatol.* 2010;24(6):769-781.
3. Hoy D, March L, Brooks P, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Ann Rheum Dis.* 2014;73(6):968-974.
4. Wu A, March L, Zheng X, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. *Ann Transl Med.* 2020;8(6):299.
5. Australian Institute of Health and Welfare. *Impacts of chronic back problems.* Canberra: AIHW;2016.
6. Dieleman JL, Cao J, Chapin A, et al. US Health Care Spending by Payer and Health Condition, 1996-2016. *JAMA.* 2020;323(9):863-884.
7. Australian Institute of Health and Welfare. *Australian Burden of Disease Study: impact and causes of illness and death in Australia 2015.* Canberra: AIHW;2019.
8. de Luca K, Briggs AM, French SD, et al. Disability burden due to musculoskeletal conditions and low back pain in Australia: findings from GBD 2019. *Chiropr Man Therap.* 2022;30(1):22.
9. Dagenais S, Caro J, Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. *Spine J.* 2008;8(1):8-20.
10. Walker BF, Muller R, Grant WD. Low back pain in Australian adults: the economic burden. *Asia Pac J Public Health.* 2003;15(2):79-87.
11. Buchbinder R, van Tulder M, Oberg B, et al. Low back pain: a call for action. *Lancet.* 2018;391(10137):2384-2388.
12. Costa N, Blyth FM, Amorim AB, Parambath S, Shanmuganathan S, Huckel Schneider C. Implementation Initiatives to Improve Low Back Pain Care in Australia: A Scoping Review. *Pain Med.* 2022;23(12):1979-2009.
13. Jacobs JC, Jarvik JG, Chou R, et al. Observational Study of the Downstream Consequences of Inappropriate MRI of the Lumbar Spine. *J Gen Intern Med.* 2020;35(12):3605-3612.
14. Traeger AC, Henschke N, Hubscher M, et al. Estimating the Risk of Chronic Pain: Development and Validation of a Prognostic Model (PICKUP) for Patients with Acute Low Back Pain. *PLoS Med.* 2016;13(5):e1002019.
15. McCaffery KJ, Jansen J, Scherer LD, et al. Walking the tightrope: communicating overdiagnosis in modern healthcare. *BMJ.* 2016;352:i348.
16. Hartvigsen J, Hancock MJ, Kongsted A, et al. What low back pain is and why we need to pay attention. *Lancet.* 2018;391(10137):2356-2367.
17. Oliveira CB, Maher CG, Pinto RZ, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *Eur Spine J.* 2018;27(11):2791-2803.
18. Downie A, Williams CM, Henschke N, et al. Red flags to screen for malignancy and fracture in patients with low back pain. *Br J Sports Med.* 2014;48(20):1518.

19. Henschke N, Maher CG, Refshauge KM, et al. Prevalence of and screening for serious spinal pathology in patients presenting to primary care settings with acute low back pain. *Arthritis Rheum.* 2009;60(10):3072-3080.
20. Williams CM, Henschke N, Maher CG, et al. Red flags to screen for vertebral fracture in patients presenting with low-back pain. *Cochrane Database Syst Rev.* 2013(1):CD008643.
21. Galliker G, Scherer DE, Trippolini MA, Rasmussen-Barr E, LoMartire R, Wertli MM. Low Back Pain in the Emergency Department: Prevalence of Serious Spinal Pathologies and Diagnostic Accuracy of Red Flags. *Am J Med.* 2020;133(1):60-72 e14.
22. Melman A, Maher CG, Needs C, Machado GC. Many people admitted to hospital with a provisional diagnosis of nonserious back pain are subsequently found to have serious pathology as the underlying cause. *Clin Rheumatol.* 2022;41(6):1867-1871.
23. Ferreira GE, Machado GC, Abdel Shaheed C, et al. Management of low back pain in Australian emergency departments. *BMJ Qual Saf.* 2019;28(10):826-834.
24. Han CS, Hancock MJ, Downie A, et al. Red flags to screen for vertebral fracture in patients presenting with low back pain. *Cochrane Database of Systematic Reviews.* 2022(7).
25. Deyo RA, Tsui-Wu YJ. Descriptive epidemiology of low-back pain and its related medical care in the United States. *Spine (Phila Pa 1976).* 1987;12(3):264-268.
26. Koes BW, van Tulder MW, Peul WC. Diagnosis and treatment of sciatica. *BMJ.* 2007;334(7607):1313-1317.
27. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *Lancet.* 2017;389(10070):736-747.
28. Knezevic NN, Candido KD, Vlaeyen JWS, Van Zundert J, Cohen SP. Low back pain. *Lancet.* 2021;398(10294):78-92.
29. Itz CJ, Willems PC, Zeilstra DJ, et al. Dutch Multidisciplinary Guideline for Invasive Treatment of Pain Syndromes of the Lumbosacral Spine. *Pain Pract.* 2016;16(1):90-110.
30. Zhao L, Manchikanti L, Kaye AD, Abd-Elsayed A. Treatment of Discogenic Low Back Pain: Current Treatment Strategies and Future Options-a Literature Review. *Curr Pain Headache Rep.* 2019;23(11):86.
31. Manchikanti L, Soin A, Benyamin RM, et al. An Update of the Systematic Appraisal of the Accuracy and Utility of Discography in Chronic Spinal Pain. *Pain Physician.* 2018;21(2):91-110.
32. Carragee EJ, Don AS, Hurwitz EL, Cuellar JM, Carrino JA, Herzog R. 2009 ISSLS Prize Winner: Does discography cause accelerated progression of degeneration changes in the lumbar disc: a ten-year matched cohort study. *Spine (Phila Pa 1976).* 2009;34(21):2338-2345.
33. DePalma MJ, Ketchum JM, Saullo T. What is the source of chronic low back pain and does age play a role? *Pain Med.* 2011;12(2):224-233.
34. Manchikanti L, Kaye AD, Boswell MV, et al. A Systematic Review and Best Evidence Synthesis of the Effectiveness of Therapeutic Facet Joint Interventions in Managing Chronic Spinal Pain. *Pain Physician.* 2015;18(4):E535-582.
35. Boswell MV, Manchikanti L, Kaye AD, et al. A Best-Evidence Systematic Appraisal of the Diagnostic Accuracy and Utility of Facet (Zygapophysial) Joint Injections in Chronic Spinal Pain. *Pain Physician.* 2015;18(4):E497-533.

36. Patel ND, Broderick DF, Burns J, et al. ACR Appropriateness Criteria Low Back Pain. *J Am Coll Radiol.* 2016;13(9):1069-1078.
37. Genant HK, Wu CY, van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. *J Bone Miner Res.* 1993;8(9):1137-1148.
38. Jiang G, Eastell R, Barrington NA, Ferrar L. Comparison of methods for the visual identification of prevalent vertebral fracture in osteoporosis. *Osteoporos Int.* 2004;15(11):887-896.
39. Ang EC, Robertson AF, Malara FA, et al. Diagnostic accuracy of 3-T magnetic resonance imaging with 3D T1 VIBE versus computer tomography in pars stress fracture of the lumbar spine. *Skeletal Radiol.* 2016;45(11):1533-1540.
40. Astur DC, Zanatta F, Arliani GG, Moraes ER, Pochini Ade C, Ejnisman B. Stress fractures: definition, diagnosis and treatment. *Rev Bras Ortop.* 2016;51(1):3-10.
41. Alqarni AM, Schneiders AG, Cook CE, Hendrick PA. Clinical tests to diagnose lumbar spondylolysis and spondylolisthesis: A systematic review. *Phys Ther Sport.* 2015;16(3):268-275.
42. Cheung KK, Dhawan RT, Wilson LF, Peirce NS, Rajeswaran G. Pars interarticularis injury in elite athletes - The role of imaging in diagnosis and management. *Eur J Radiol.* 2018;108:28-42.
43. Murthy NS. Imaging of stress fractures of the spine. *Radiol Clin North Am.* 2012;50(4):799-821.
44. Verhagen AP, Downie A, Popal N, Maher C, Koes BW. Red flags presented in current low back pain guidelines: a review. *Eur Spine J.* 2016;25(9):2788-2802.
45. Finucane LM, Downie A, Mercer C, et al. International Framework for Red Flags for Potential Serious Spinal Pathologies. *J Orthop Sports Phys Ther.* 2020;50(7):350-372.
46. Henschke N, Maher CG, Ostelo RW, de Vet HC, Macaskill P, Irwig L. Red flags to screen for malignancy in patients with low-back pain. *Cochrane Database Syst Rev.* 2013(2):CD008686.
47. Dionne N, Adefolarin A, Kunzelman D, et al. What is the diagnostic accuracy of red flags related to cauda equina syndrome (CES), when compared to Magnetic Resonance Imaging (MRI)? A systematic review. *Musculoskelet Sci Pract.* 2019;42:125-133.
48. Han CS, Hancock MJ, Maher CG. Reconsidering non-specific low back pain: where to from here? *Spine J.* 2022;22(12):1927-1930.
49. Hancock MJ, Maher CG, Latimer J, et al. Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. *Eur Spine J.* 2007;16(10):1539-1550.
50. Hayden JA, Ellis J, Ogilvie R, Malmivaara A, van Tulder MW. Exercise therapy for chronic low back pain. *Cochrane Database Syst Rev.* 2021;9(9):CD009790.
51. Ho EK, Chen L, Simic M, et al. Psychological interventions for chronic, non-specific low back pain: systematic review with network meta-analysis. *BMJ.* 2022;376:e067718.
52. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess.* 2007;11(50):iii, ix-51.
53. International Organization for the Study of Pain. Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms. In:2011: https://www.iasp-

pain.org/publications/free-ebooks/classification-of-chronic-pain-second-edition-revised/. Accessed 2022.
54. Derby R, Kim BJ, Lee SH, Chen Y, Seo KS, Aprill C. Comparison of discographic findings in asymptomatic subject discs and the negative discs of chronic LBP patients: can discography distinguish asymptomatic discs among morphologically abnormal discs? *Spine J.* 2005;5(4):389-394.
55. Walsh TR, Weinstein JN, Spratt KF, Lehmann TR, Aprill C, Sayre H. Lumbar discography in normal subjects. A controlled, prospective study. *J Bone Joint Surg Am.* 1990;72(7):1081-1088.
56. Brinjikji W, Luetmer PH, Comstock B, et al. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *AJNR Am J Neuroradiol.* 2015;36(4):811-816.
57. Apeldoorn AT, van Helvoirt H, Meihuizen H, et al. The Influence of Centralization and Directional Preference on Spinal Control in Patients With Nonspecific Low Back Pain. *J Orthop Sports Phys Ther.* 2016;46(4):258-269.
58. Saueressig T, Owen PJ, Diemer F, Zebisch J, Belavy DL. Diagnostic Accuracy of Clusters of Pain Provocation Tests for Detecting Sacroiliac Joint Pain: Systematic Review With Meta-analysis. *J Orthop Sports Phys Ther.* 2021;51(9):422-431.
59. Jarvik JG, Hollingworth W, Martin B, et al. Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomized controlled trial. *JAMA.* 2003;289(21):2810-2818.
60. Deyo RA, Mirza SK, Turner JA, Martin BI. Overtreating chronic back pain: time to back off? *J Am Board Fam Med.* 2009;22(1):62-68.
61. Koes BW, van Tulder MW, Thomas S. Diagnosis and treatment of low back pain. *BMJ.* 2006;332(7555):1430-1434.
62. Albert HB, Kjaer P, Jensen TS, Sorensen JS, Bendix T, Manniche C. Modic changes, possible causes and relation to low back pain. *Med Hypotheses.* 2008;70(2):361-368.
63. Albert HB, Sorensen JS, Christensen BS, Manniche C. Antibiotic treatment in patients with chronic low back pain and vertebral bone edema (Modic type 1 changes): a double-blind randomized clinical controlled trial of efficacy. *Eur Spine J.* 2013;22(4):697-707.
64. Cai G, Laslett LL, Aitken D, et al. Effect of Zoledronic Acid and Denosumab in Patients With Low Back Pain and Modic Change: A Proof-of-Principle Trial. *J Bone Miner Res.* 2018;33(5):773-782.
65. Steffens D, Hancock MJ, Maher CG, Williams C, Jensen TS, Latimer J. Does magnetic resonance imaging predict future low back pain? A systematic review. *Eur J Pain.* 2014;18(6):755-765.
66. Jang HD, Kim EH, Lee JC, Choi SW, Kim K, Shin BJ. Current Concepts in the Management of Osteoporotic Vertebral Fractures: A Narrative Review. *Asian Spine J.* 2020;14(6):898-909.
67. McCarthy J, Davis A. Diagnosis and Management of Vertebral Compression Fractures. *Am Fam Physician.* 2016;94(1):44-50.
68. Parreira PCS, Maher CG, Megale RZ, March L, Ferreira ML. An overview of clinical guidelines for the management of vertebral compression fracture: a systematic review. *Spine J.* 2017;17(12):1932-1938.

69. Lee HM, Park SY, Lee SH, Suh SW, Hong JY. Comparative analysis of clinical outcomes in patients with osteoporotic vertebral compression fractures (OVCFs): conservative treatment versus balloon kyphoplasty. *Spine J.* 2012;12(11):998-1005.
70. Barr JD, Jensen ME, Hirsch JA, et al. Position statement on percutaneous vertebral augmentation: a consensus statement developed by the Society of Interventional Radiology (SIR), American Association of Neurological Surgeons (AANS) and the Congress of Neurological Surgeons (CNS), American College of Radiology (ACR), American Society of Neuroradiology (ASNR), American Society of Spine Radiology (ASSR), Canadian Interventional Radiology Association (CIRA), and the Society of NeuroInterventional Surgery (SNIS). *J Vasc Interv Radiol.* 2014;25(2):171-181.
71. McConnell CT, Jr., Wippold FJ, 2nd, Ray CE, Jr., et al. ACR appropriateness criteria management of vertebral compression fractures. *J Am Coll Radiol.* 2014;11(8):757-763.
72. Savage JW, Schroeder GD, Anderson PA. Vertebroplasty and kyphoplasty for the treatment of osteoporotic vertebral compression fractures. *J Am Acad Orthop Surg.* 2014;22(10):653-664.
73. Venmans A, Lohle PN, van Rooij WJ. Pain course in conservatively treated patients with back pain and a VCF on the spine radiograph (VERTOS III). *Skeletal Radiol.* 2014;43(1):13-18.
74. Gray LA, Jarvik JG, Heagerty PJ, et al. INvestigational Vertebroplasty Efficacy and Safety Trial (INVEST): a randomized controlled trial of percutaneous vertebroplasty. *BMC Musculoskelet Disord.* 2007;8:126.
75. Hussain A, Erdek M. Vertebroplasty augmentation procedures: examining the controversy. *Pain Physician.* 2013;16(5):E483-490.
76. Liu JT, Liao WJ, Tan WC, et al. Balloon kyphoplasty versus vertebroplasty for treatment of osteoporotic vertebral compression fracture: a prospective, comparative, and randomized clinical study. *Osteoporos Int.* 2010;21(2):359-364.
77. Rousing R, Andersen MO, Jespersen SM, Thomsen K, Lauritsen J. Percutaneous vertebroplasty compared to conservative treatment in patients with painful acute or subacute osteoporotic vertebral fractures: three-months follow-up in a clinical randomized study. *Spine (Phila Pa 1976).* 2009;34(13):1349-1354.
78. Esses SI, McGuire R, Jenkins J, et al. American Academy of Orthopaedic Surgeons clinical practice guideline on: the treatment of osteoporotic spinal compression fractures. *J Bone Joint Surg Am.* 2011;93(20):1934-1936.
79. Koes BW, van Tulder M, Lin CW, Macedo LG, McAuley J, Maher C. An updated overview of clinical guidelines for the management of non-specific low back pain in primary care. *Eur Spine J.* 2010;19(12):2075-2094.
80. Abdel Shaheed C, Ferreira GE, Dmitritchenko A, et al. The efficacy and safety of paracetamol for pain relief: an overview of systematic reviews. *Med J Aust.* 2021;214(7):324-331.
81. Jones CMP, Lin CC, Jamshidi M, et al. Effectiveness of Opioid Analgesic Medicines Prescribed in or at Discharge From Emergency Departments for Musculoskeletal Pain : A Systematic Review and Meta-analysis. *Ann Intern Med.* 2022;175(11):1572-1581.
82. van der Gaag WH, Roelofs PD, Enthoven WT, van Tulder MW, Koes BW. Non-steroidal anti-inflammatory drugs for acute low back pain. *Cochrane Database Syst Rev.* 2020;4(4):CD013581.

83. Williams CM, Maher CG, Latimer J, et al. Efficacy of paracetamol for acute low-back pain: a double-blind, randomised controlled trial. *Lancet.* 2014;384(9954):1586-1596.
84. National Institute for Health and Care Excellence. *Non-specific low back pain and sciatica: management. NICE guideline: short version. Draft for consultation.* March 2016 2016.
85. Kent P, Haines T, O'Sullivan P, et al. Cognitive functional therapy with or without movement sensor biofeedback versus usual care for chronic, disabling low back pain (RESTORE): a randomised, controlled, three-arm, parallel group, phase 3, clinical trial. *Lancet.* 2023.
86. Kamper SJ, Apeldoorn AT, Chiarotto A, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: Cochrane systematic review and meta-analysis. *BMJ.* 2015;350:h444.
87. Vibe Fersum K, O'Sullivan P, Skouen JS, Smith A, Kvale A. Efficacy of classification-based cognitive functional therapy in patients with non-specific chronic low back pain: a randomized controlled trial. *Eur J Pain.* 2013;17(6):916-928.
88. O'Keeffe M, O'Sullivan P, Purtill H, Bargary N, O'Sullivan K. Cognitive functional therapy compared with a group-based exercise and education intervention for chronic low back pain: a multicentre randomised controlled trial (RCT). *Br J Sports Med.* 2020;54(13):782-789.
89. Nury E, Schmucker C, Nagavci B, et al. Efficacy and safety of strong opioids for chronic noncancer pain and chronic low back pain: a systematic review and meta-analyses. *Pain.* 2022;163(4):610-636.
90. Ferreira GE, Abdel-Shaheed C, Underwood M, et al. Efficacy, safety, and tolerability of antidepressants for pain in adults: overview of systematic reviews. *BMJ.* 2023;380:e072415.
91. Enke O, New HA, New CH, et al. Anticonvulsants in the treatment of low back pain and lumbar radicular pain: a systematic review and meta-analysis. *CMAJ.* 2018;190(26):E786-E793.
92. Chou R, Qaseem A, Snow V, et al. 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.* 2007;147(7):478-491.
93. Chowdhury AR, Graham PL, Schofield D, Cunich M, Nicholas M. Cost-effectiveness of Multidisciplinary Interventions for Chronic Low Back Pain: A Narrative Review. *Clin J Pain.* 2021;38(3):197-207.
94. Foster NE, Mullis R, Hill JC, et al. Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. *Ann Fam Med.* 2014;12(2):102-111.
95. Hill JC, Whitehurst DG, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet.* 2011;378(9802):1560-1571.
96. Cherkin D, Balderson B, Wellman R, et al. Effect of Low Back Pain Risk-Stratification Strategy on Patient Outcomes and Care Processes: the MATCH Randomized Trial in Primary Care. *J Gen Intern Med.* 2018;33(8):1324-1336.

97. Konstantinou K, Lewis M, Dunn KM, et al. Stratified care versus usual care for management of patients presenting with sciatica in primary care (SCOPiC): a randomised controlled trial. *Lancet Rheumatol.* 2020;2(7):e401-e411.
98. Koppenaal T, Pisters MF, Kloek CJ, Arensman RM, Ostelo RW, Veenhof C. The 3-Month Effectiveness of a Stratified Blended Physiotherapy Intervention in Patients With Nonspecific Low Back Pain: Cluster Randomized Controlled Trial. *J Med Internet Res.* 2022;24(2):e31675.
99. Morso L, Olsen Rose K, Schiottz-Christensen B, Sowden G, Sondergaard J, Christiansen DH. Effectiveness of stratified treatment for back pain in Danish primary care: A randomized controlled trial. *Eur J Pain.* 2021;25(9):2020-2038.
100. Beneciuk JM, George SZ. Pragmatic Implementation of a Stratified Primary Care Model for Low Back Pain Management in Outpatient Physical Therapy Settings: Two-Phase, Sequential Preliminary Study. *Phys Ther.* 2015;95(8):1120-1134.
101. Choudhry NK, Fifer S, Fontanet CP, et al. Effect of a Biopsychosocial Intervention or Postural Therapy on Disability and Health Care Spending Among Patients With Acute and Subacute Spine Pain: The SPINE CARE Randomized Clinical Trial. *JAMA.* 2022;328(23):2334-2344.
102. Delitto A, Patterson CG, Stevans JM, et al. Stratified care to prevent chronic low back pain in high-risk patients: The TARGET trial. A multi-site pragmatic cluster randomized trial. *EClinicalMedicine.* 2021;34:100795.
103. Murphy SE, Blake C, Power CK, Fullen BM. Comparison of a Stratified Group Intervention (STarT Back) With Usual Group Care in Patients With Low Back Pain: A Nonrandomized Controlled Trial. *Spine (Phila Pa 1976).* 2016;41(8):645-652.
104. Rhon DI, Greenlee TA, Poehlein E, et al. Effect of Risk-Stratified Care on Disability Among Adults With Low Back Pain Treated in the Military Health System: A Randomized Clinical Trial. *JAMA Netw Open.* 2023;6(7):e2321929.
105. Tagliaferri SD, Mitchell UH, Saueressig T, Owen PJ, Miller CT, Belavy DL. Classification Approaches for Treating Low Back Pain Have Small Effects That Are Not Clinically Meaningful: A Systematic Review With Meta-analysis. *J Orthop Sports Phys Ther.* 2022;52(2):67-84.
106. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ.* 2010;340:c117.
107. Braten LCH, Rolfsen MP, Espeland A, et al. Efficacy of antibiotic treatment in patients with chronic low back pain and Modic changes (the AIM study): double blind, randomised, placebo controlled, multicentre trial. *BMJ.* 2019;367:l5654.
108. Weiler S, Alyagon B. Phase 1 Study of Antibiotic Treatment for Low Back Pain.
109. Shea GKH, Zhang C, Suen WS, et al. Oral Zoledronic acid bisphosphonate for the treatment of chronic low back pain with associated Modic changes: A pilot randomized controlled trial. *J Orthop Res.* 2022;40(12):2924-2936.
110. Lee SW, Han HC. Methylene Blue Application to Lessen Pain: Its Analgesic Effect and Mechanism. *Front Neurosci.* 2021;15:663650.
111. Peng B, Pang X, Wu Y, Zhao C, Song X. A randomized placebo-controlled trial of intradiscal methylene blue injection for the treatment of chronic discogenic low back pain. *Pain.* 2010;149(1):124-129.
112. Peng B, Zhang Y, Hou S, Wu W, Fu X. Intradiscal methylene blue injection for the treatment of chronic discogenic low back pain. *Eur Spine J.* 2007;16(1):33-38.

113. Kallewaard JW, Wintraecken VM, Geurts JW, et al. A multicenter randomized controlled trial on the efficacy of intradiscal methylene blue injection for chronic discogenic low back pain: the IMBI study. *Pain.* 2019;160(4):945-953.
114. Hansen H, Manchikanti L, Simopoulos TT, et al. A systematic evaluation of the therapeutic effectiveness of sacroiliac joint interventions. *Pain Physician.* 2012;15(3):E247-278.
115. Manchikanti L, Pampati V, Falco FJ, Hirsch JA. An updated assessment of utilization of interventional pain management techniques in the Medicare population: 2000 - 2013. *Pain Physician.* 2015;18(2):E115-127.
116. Kennedy DJ, Engel A, Kreiner DS, Nampiaparampil D, Duszynski B, MacVicar J. Fluoroscopically Guided Diagnostic and Therapeutic Intra-Articular Sacroiliac Joint Injections: A Systematic Review. *Pain Med.* 2015;16(8):1500-1518.
117. Zheng P, Schneider BJ, Yang A, McCormick ZL. Image-Guided Sacroiliac Joint Injections: an Evidence-based Review of Best Practices and Clinical Outcomes. *PM R.* 2019;11 Suppl 1:S98-S104.
118. Leggett LE, Soril LJ, Lorenzetti DL, et al. Radiofrequency ablation for chronic low back pain: a systematic review of randomized controlled trials. *Pain Res Manag.* 2014;19(5):e146-153.
119. Juch JNS, Maas ET, Ostelo R, et al. Effect of Radiofrequency Denervation on Pain Intensity Among Patients With Chronic Low Back Pain: The Mint Randomized Clinical Trials. *JAMA.* 2017;318(1):68-81.
120. Maas ET, Ostelo RW, Niemisto L, et al. Radiofrequency denervation for chronic low back pain. *Cochrane Database Syst Rev.* 2015;2015(10):CD008572.
121. Deyo RA. Diagnostic evaluation of LBP: reaching a specific diagnosis is often impossible. *Arch Intern Med.* 2002;162(13):1444-1447; discussion 1447-1448.
122. Lequerre T, Rottenberg P, Derambure C, Cosette P, Vittecoq O. Predictors of treatment response in rheumatoid arthritis. *Joint Bone Spine.* 2019;86(2):151-158.
123. Smith SL, Plant D, Eyre S, Barton A. The potential use of expression profiling: implications for predicting treatment response in rheumatoid arthritis. *Ann Rheum Dis.* 2013;72(7):1118-1124.
124. Viatte S, Barton A. Genetics of rheumatoid arthritis susceptibility, severity, and treatment response. *Semin Immunopathol.* 2017;39(4):395-408.
125. Herlin C, Kjaer P, Espeland A, et al. Modic changes-Their associations with low back pain and activity limitation: A systematic literature review and meta-analysis. *PLoS One.* 2018;13(8):e0200677.
126. Han CS, Maher CG, Steffens D, et al. Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review. *J Physiother.* 2023.

Chapter Two:

Reconsidering non-specific low back pain: where to from here?

Statement from co-authors confirming authorship contribution of the PhD candidate

The co-authors of the paper *"Han CS, Hancock MJ, Maher CG. Reconsidering non-specific low back pain: where to from here? Spine J. 2022 Dec;22(12):1927-1930. doi: 10.1016/j.spinee.2022.08.001. Epub 2022 Aug 6. PMID: 35944828."* confirm that Christopher S Han has provided the following contributions to the study:

- Conception and design of the research
- Data acquisition
- Data analysis and interpretation of findings
- Writing of the manuscript and critical appraisal of the content

Christopher S Han _____ Date: 22/11/2023

Professor Chris G Maher _____ Date: 22/11/2023

Professor Mark J Hancock _____ Date: 22/11/2023

The Spine Journal 22 (2022) 1927–1930

Perspective

Reconsidering non-specific low back pain: where to from here?

Christopher S. Han, BAppSci (OT), DPT[a,*],
Mark J. Hancock, BAppSci (Phty), MAppSc, PhD[b],
Christopher G. Maher, BAppSc (Phty), PhD, DMedSc, FACP, FAAHMS[a]

[a] *The Institute for Musculoskeletal Health, The University of Sydney and Sydney Local Health District, PO Box M179, Missenden Rd, NSW 2050, Sydney, Australia*
[b] *The Faculty of Medicine, Health and Human Sciences Level 3, 75 Talavera Rd, Macquarie University, NSW 2109, Australia*
Received 28 July 2022; revised 2 August 2022; accepted 2 August 2022

Keywords: Casual; mechanisms; LBP; nociceptive; source; non-specific

The term non-specific low back pain (NSLBP) entered medical literature at least six decades ago [1], but seems misunderstood by some. Two key misunderstandings are: using the term NSLBP automatically results in non-specific treatment; and diagnostic research seeking to identify specific causes (eg, nociceptive sources and causal mechanisms) of low back pain (LBP) is of no value. Our perspective considers how these misconceptions may have arisen, explains how these views can impede progress in science and delivery of quality clinical care, and suggests directions for future research.

Misunderstanding the intentions of those who coined the term NSLBP

In 1956 Shaw and Taylor used the term NSLBP to describe patients where a specific diagnosis was not possible even after a full clinical examination supplemented by imaging and laboratory tests [1]. In 1963 Rowe and Rochester noted that a definitive diagnosis for LBP was often unable to be established [2]. However, the proponents of the term were not advocating that each patient with NSLBP should receive the same treatment. Leaders in the field argued for the need to individualize care for people with NSLBP. For example, in 1981 McKenzie proposed a classification system to individualize care for LBP [3]. In 1987 the orthopedic surgeon Waddell promoted the bio-psychosocial model as an "approach to both the physical and the psychosocial problems of each individual patient" to assist in delivering individualized care in people with NSLBP [4]. Waddell, however, also pointed to the limitations of the term NSLBP stating the term "is intellectually and scientifically inadequate and fails to provide any biological basis for real understanding" [5].

In more recent times some commentators have argued against the use of the term NSLBP on the grounds that it leads to non-specific treatment [5,6]. While this position ignores the position of those who originally coined the term, there is some evidence to support their concerns. For example, many current clinical practice guidelines recommend broad treatment approaches such as exercise, medication, or manual therapy for patients with NSLBP [7], but provide very little to no information and guidance on how to choose between these options and how to individualize these treatments for patients with NSLBP (eg, type or dose of exercise).

Misunderstanding NSLBP may have impeded important research

It is possible that misunderstandings about NSLBP have influenced the types of LBP research that have occurred. Our view is that research investigating generic interventions for NSLBP is more common than research evaluating individualized care of NSLBP and causes of LBP. The original proponents used the term NSLBP to highlight the current diagnostic challenge, but did not suggest it would remain impossible for all time to identify causes of LBP or

FDA device/drug status: Not applicable.

Author disclosures: ***CSH***: Nothing to disclose. ***MJH***: Nothing to disclose. ***CGM***: Nothing to disclose.

*Corresponding author. The Institute for Musculoskeletal Health, The University of Sydney and Sydney Local Health District, PO Box M179, Missenden Rd, NSW 2050, Sydney, Australia. Tel.: 02 8627 7423.
E-mail address: christopher.han@sydney.edu.au (C.S. Han).

https://doi.org/10.1016/j.spinee.2022.08.001

Diagnosis	Non-specific LBP			Specific LBP
Approach	Match patient to a specific treatment	Consider patient preference & clinician experience	Consider causal mediators	Identify new categories of specific LBP currently bundled within NSLBP category
Example	*Treatment-based stratification e.g., StarTBacK, MDT, TBC*	*Shared decision making, e.g., EBM model*	*Treatments (e.g., CBT) targeting individual mediators of pain e.g., sleep disturbance, psychological distress, and central sensitisation*	*Develop effective treatments that address a specific structure and pathology e.g., annular fissure, disc infection, facet joint OA*

Fig. 1. Approaches to direct future research towards individualizing care to move beyond generic treatments for NSLBP.

individualize treatment [4], nor did they suggest that research into these areas should not be undertaken. For instance, Waddell's seminal paper suggested that future research should "develop a rational basis for choosing the most effective treatment for individual patients [4]."

Some research effort has been made to individualize care for people with NSLBP or identify subgroups; however, most of this work has focused on matching patients with treatments based on prognostic or other patient factors without considering causes or mechanisms. Stratification systems (eg, McKenzie, Treatment Based Classification, STarT Back Tool) initially demonstrated small positive effects, however a recent review found insufficient evidence to support the use of these systems compared to non-specific interventions for NSLBP [8]. There has been limited research aiming to identify the causes and mechanisms of a patient's LBP, which has limited the development and testing of specific treatments targeting these causes and mechanisms.

Current ability to identify specific nociceptive sources of pain

The current evidence suggests that we lack clinical tests to specify the nociceptive source (eg, Modic type 1, annular fissure, facet joint degeneration) responsible for a patient's NSLBP [9], however it needs to be recognized that little research has been conducted in this field. For example Hancock et al.'s 2007 [6] systematic review found only 28 studies that investigated the disc as source of LBP, 8 studies investigated the facet joint and 7 studies investigated the SIJ. Few index tests were investigated in more than one study and study quality was generally poor. This review [9] is now over 15 years old and to our knowledge only 2 reviews investigating the sacroiliac joint and only 1 review investigating the disc as the nociceptive source of LBP have been published since, with a limited number of additional primary studies.

The diagnostic usefulness of tests to identify the nociceptive source of LBP in clinical practice, particularly to guide specific treatment selection and improve treatment outcomes, currently remains unclear [10]. For example, the review by Steffens et al. 2016 identified only eight low quality trials that investigated MRI findings as potential modifiers of the effect of specific interventions [10]. In no study was this primary aim or was the intervention a conservative intervention [10]. Some research has attempted to identify the pathoanatomical causes and mechanisms (eg, infection, trauma, genetics) of LBP with the intent to develop specific treatments targeting such causes and mechanisms. For example, Modic Type 1 changes are a potential nociceptive source of LBP and efforts have been made to identify the pathology and causal mechanisms (eg, bacterial infection, inflammation, and bone marrow oedema) [11]. Treatments that aim to address these causes and mechanisms (eg, antibiotics, Zoledronic acid, Denosumab) require much further research regarding their safety and effectiveness; however, these research approaches present an opportunity to develop treatments that target a specific cause of NSLBP.

Future directions for LBP research

There are several research approaches to help us move beyond non-specific interventions for NSLBP and hopefully improve patient's outcomes. In Fig. 1 we present four possible approaches for individualizing care for patients with LBP. Our view is that we need a balanced portfolio of LBP research; however, we believe there should be greater attention on research investigating the pathoanatomical basis of LBP (eg, Modic type 1, annular fissure, facet joint degeneration) and identifying the causes and mechanisms (eg, infection, trauma, genetics) of LBP. Advancements in medical technology and imaging in the future may present new research opportunities to investigate the pathoanatomical basis of LBP. A greater understanding of the causes (nociceptive sources and causal mechanisms) will likely lead to new management approaches that directly target the specific nociceptive source and/or causal mechanisms contributing to LBP and its consequent disability and help move past generic interventions. For example, rheumatoid arthritis was once a debilitating condition leading to deformities and joint damage, however, advancements in understanding pathological mechanisms behind rheumatoid arthritis have led to the development of more targeted treatments. Simultaneously, we should also continue investigating the effects of refining treatments based on causal mediators. Recent studies have identified causal mediators for individuals with NSLBP (eg, sleep disturbance, psychological distress, and central sensitization), which are associated with greater levels of pain [12]. These causal mediators can be used as targets to refine treatments specific to the patient (eg, cognitive behavioral therapy). Future high-quality studies identifying groups of patients with common causal mediators may provide promising opportunities to move away from generic treatments by refining existing treatments (or develop new treatments) and better match patients to specific treatments. Further research is needed to determine if

these approaches will lead to more individualized treatments and better outcomes.

How can clinicians currently individualize care if they cannot accurately identify the specific cause of a patient's pain?

There are some approaches that may help clinicians individualize treatment of NSLBP. The evidence-based practice model has always suggested that clinicians individualize care by matching the patient to treatments through consideration of the values and preferences of the patient and clinical expertise of the clinician. Consideration of patient preferences in a shared decision approach provides one easy way for clinicians to move beyond generic treatments and instead offer personalized care [13] for NSLBP. Systematic reviews have shown shared decision-making improves patient engagement and may improve patient outcomes [14].

For a proportion of patients their LBP may not be explained by a pathoanatomical change in a peripheral structure and may instead be better explained by central sensitization. Patients classified as having central sensitization, will likely require specific treatments targeting the mechanisms underlying the hyperexcitability of the central nervous system [15]. Treatments aim to desensitize the central nervous system through approaches such as pharmacological options, electrotherapy targeting the brain, exercise therapy, cognitive behavioral therapy, and pain neuroscience education [15]. However more research is needed to support these treatment approaches for patients presenting with primary central sensitization.

Clinicians can also individualize care by refining treatments and implementing treatments based on the presenting clinical features of the patient [16,17]. At present there is some evidence that personalizing care in this way has merit [17], but requires more investigation. A recent individual patient data meta-analysis (n=19 trials; 9,328 participants) [18] found those with higher levels of disability were more likely to benefit from psychological treatments. Those who were younger, had higher levels of disability, and low levels of psychological distress were more likely to benefit from passive physical treatments. However, other IPD meta-analyses investigating acupuncture (n=39 trials; 20,827 participants) [19] and spinal manipulative therapy (n=43 trials; 4,223 participants) [20] for NSLBP did not identify patient characteristics that identified individuals that would gain greater benefit from those treatments.

Conclusion

Non-specific low back pain is a term we should use now, but hopefully not forever, or at least for a smaller proportion of patients as research progresses. The common misunderstandings about this term have held back progress in science and clinical practice for LBP. We argue that we need a balanced portfolio of LBP research and that research investigating the pathoanatomical basis and causes and mechanisms of LBP should be part of this portfolio.

Funding source

None.

Ethics and dissemination

Ethical approval is not required.

Financial Disclosure and Conflict of Interest

I affirm that I have no financial affiliation (including research funding) or involvement with any commercial organization that has a direct financial interest in any matter included in this manuscript, except as disclosed in an attachment and cited in the manuscript. Any other conflict of interest (ie, personal associations or involvement as a director, officer, or expert witness) is also disclosed in an attachment.

Declarations of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

None.

References

[1] Shaw EG, Taylor JG. The results of lumbo-sacral fusion for low back pain. J Bone Joint Surg Br 1956;38-B(2):485–97.
[2] Rowe ML. Preliminary statistical study of low back pain. J Occup Med 1963;5:336–41.
[3] Santolin SM. McKenzie diagnosis and therapy in the evaluation and management of a lumbar disc derangement syndrome: a case study. J Chiropr Med 2003;2(2):60–5.
[4] Waddell G. Volvo award in clinical sciences. A new clinical model for the treatment of low-back pain. Spine (Phila Pa 1976) 1987;12 (7):632–44.
[5] Waddell G. Subgroups within "nonspecific" low back pain. J Rheumatol 2005;32(3):395–6.
[6] Petersen T, Laslett M, Juhl C. Clinical classification in low back pain: best-evidence diagnostic rules based on systematic reviews. BMC Musculoskelet Disord 2017;18(1):188.
[7] Oliveira CB, Maher CG, Pinto RZ, Traeger AC, Lin CC, Chenot JF, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. Eur Spine J 2018;27(11):2791–803.
[8] Tagliaferri SD, Mitchell UH, Saueressig T, Owen PJ, Miller CT, Belavy DL. Classification approaches for treating low back pain have small effects that are not clinically meaningful: a systematic review with meta-analysis. J Orthop Sports Phys Ther 2022;52(2):67–84.
[9] Hancock MJ, Maher CG, Latimer J, Spindler MF, McAuley JH, Laslett M, et al. Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. Eur Spine J 2007;16(10):1539–50.

[10] Steffens D, Hancock MJ, Pereira LS, Kent PM, Latimer J, Maher CG. Do MRI findings identify patients with low back pain or sciatica who respond better to particular interventions? A systematic review. Eur Spine J 2016;25(4):1170–87.
[11] Cai G, Laslett LL, Aitken D, Halliday A, Pan F, Otahal P, et al. Effect of zoledronic acid and Denosumab in patients with low back pain and modicchange: a proof-of-principle trial. J Bone Miner Res 2018;33 (5):773–82.
[12] Carlesso LC, Tousignant-Laflamme Y, Shaw W, Lariviere C, Choiniere M. Exploring pain phenotypes in workers with chronic low back pain: application of IMMPACT recommendations. Can J Pain 2021;5(1):43–55.
[13] Herbert RDSC, Maher CG, Moseley AM. Evidence-based practice: imperfect but necessary. Physiother Theory Pract 2001;17:201–11.
[14] Durand MA, Carpenter L, Dolan H, Bravo P, Mann M, Bunn F, et al. Do interventions designed to support shared decision-making reduce health inequalities? A systematic review and meta-analysis. PLoS One 2014;9(4):e94670.
[15] Nijs J, Apeldoorn A, Hallegraeff H, Clark J, Smeets R, Malfliet A, et al. Low back pain: guidelines for the clinical classification of predominant neuropathic, nociceptive, or central sensitization pain. Pain Physician 2015;18(3):E333–46.
[16] Kent P, Keating J. Do primary-care clinicians think that nonspecific low back pain is one condition? Spine (Phila Pa 1976) 2004;29 (9):1022–31.
[17] O'Keeffe M, O'Sullivan P, Purtill H, Bargary N, O'Sullivan K. Cognitive functional therapy compared with a group-based exercise and education intervention for chronic low back pain: a multicenter randomized controlled trial (RCT). Br J Sports Med 2020;54(13):782–9.
[18] Hee SW, Mistry D, Friede T, Lamb SE, Stallard N, Underwood M, Patel S, et al. Identification of subgroup effect with an individual participant data meta-analysis of randomized controlled trials of three different types of therapist-delivered care in low back pain. BMC Musculoskelet Disord 2021;22(1):191.
[19] Foster NE, Vertosick EA, Lewith G, Linde K, MacPherson H, Sherman KJ, et al. Identifying patients with chronic pain who respond to acupuncture: results from an individual patient data meta-analysis. Acupunct Med 2021;39(2):83–90.
[20] de Zoete A, de Boer MR, Rubinstein SM, van Tulder MW, Underwood M, Hayden JA. Moderators of the effect of spinal manipulative therapy on pain relief and function in patients with chronic low back pain: an individual participant data meta-analysis. Spine (Phila Pa 1976) 2021;46(8):E505–17.

Chapter Three:

Red flags to screen for vertebral fracture in patients presenting with low back pain (Protocol)

Statement from co-authors confirming authorship contribution of the PhD candidate

The co-authors of the paper *"Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Maher CG. Red flags to screen for vertebral fracture in patients presenting with low back pain. Cochrane Database Syst Rev. 2022 Jul 20;2022(7):CD014461. doi: 10.1002/14651858.CD014461. PMCID: PMC9298669."* confirm that Christopher S Han has provided the following contributions to the study:

- Conception and design of the research
- Data acquisition
- Data analysis and interpretation of findings
- Writing of the manuscript and critical appraisal of the content

Christopher S Han _____ Date: 22/11/2023

Professor Chris G Maher _____ Date: 22/11/2023

Professor Mark J Hancock _____ Date: 22/11/2023

Cochrane Library

Cochrane Database of Systematic Reviews

Red flags to screen for vertebral fracture in patients presenting with low back pain (Protocol)

Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Maher CG

Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Maher CG.
Red flags to screen for vertebral fracture in patients presenting with low back pain (Protocol).
Cochrane Database of Systematic Reviews 2022, Issue 7. Art. No.: CD014461.
DOI: 10.1002/14651858.CD014461.

www.cochranelibrary.com

TABLE OF CONTENTS

[Diagnostic Test Accuracy Protocol]

Red flags to screen for vertebral fracture in patients presenting with low back pain

Christopher S Han[1], Mark J Hancock[2], Aron Downie[1,2], Jeffrey G Jarvik[3], Bart W Koes[4], Gustavo C Machado[1], Arianne P Verhagen[5], Christopher M Williams[6], Christopher G Maher[1]

[1]Institute for Musculoskeletal Health, The University of Sydney and Sydney Local Health District, Sydney, Australia. [2]Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia. [3]Radiology and Neurological Surgery, Clinical Learning, Evidence And Research (CLEAR) Center, University of Washington School of Medicine, Seattle, USA. [4]Center for Muscle and Health, University of Southern Denmark, Odense , Denmark. [5]Discipline of Physiotherapy, Graduate School of Health, University of Technology , Sydney, Australia. [6]School of Medicine and Public Health, University of Newcastle, Callaghan, Australia

Contact: Christopher S Han, christopher.han@sydney.edu.au.

Editorial group: Cochrane Back and Neck Group.
Publication status and date: New, published in Issue 7, 2022.

Citation: Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Maher CG. Red flags to screen for vertebral fracture in patients presenting with low back pain (Protocol). *Cochrane Database of Systematic Reviews* 2022, Issue 7. Art. No.: CD014461. DOI: 10.1002/14651858.CD014461.

ABSTRACT

Objectives

This is a protocol for a Cochrane Review (diagnostic). The objectives are as follows:

The objective of this review is to assess the diagnostic accuracy of red flags used to screen for vertebral fracture in patients presenting with low back pain. Where possible, we will report the results of red flags separately for different types of vertebral fracture as follows.

- Osteoporotic vertebral compression fracture
- Vertebral traumatic fracture
- Vertebral stress fracture

Secondary objectives

The secondary objective of this review is to assess the influence of sources of heterogeneity on the diagnostic accuracy of red flag tests, if appropriate. In particular, we will assess the healthcare setting (e.g. primary, emergency, or secondary care) and the different types of vertebral fracture (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, and vertebral stress fracture. We will also assess mean age at different cut-off thresholds (e.g. > 50 years of age versus > 70 years of age) across different studies, as this has previously been shown to influence results.

BACKGROUND

Low back pain continues to be the leading cause for years lived with disability according to The Global Burden of Disease Study (Wu 2020). It is estimated that 577 million people experience low back pain globally, and in 2017 the estimated global years lived with disability was 64.9 million (Wu 2020). It is widely agreed that low back pain can be very disabling and impose an enormous social and economic burden on the community (Buchbinder 2018). Despite the disability burden of low back pain, most presentations are not due to a serious underlying condition (Koes 2006), and can be managed in primary care without further diagnostic investigation (Oliveira 2018). However, there is an increasing number of low back pain presentations to emergency departments for care (Ferreira 2021).

Guidelines recommend that a thorough history and clinical assessment can be used to identify patients with suspicion of serious pathology (e.g. cauda equina syndrome, spinal tumours, cancer) who then require further diagnostic work-up to confirm the diagnosis. In the primary care setting, between 1% and 5% of all patients who present with low back pain will have a serious spinal pathology (Downie 2013; Hartvigsen 2018; Henschke 2009; Williams 2013). The prevalence of serious pathology in patients presenting to the emergency department is thought to be somewhat higher, one review reporting between 2.5% to 5.1% (Galliker 2019). In both settings vertebral fractures are amongst the most common serious pathologies which a clinician needs to screen for.

The use of red flags with high false-positive rates may contribute to unnecessary imaging and costs in low back pain management (Williams 2013). On the other hand, red flags with high false-negative rates can lead to failure to recognise serious pathology, resulting in delayed testing and treatment, and may increase patient morbidity (DePalma 2020).

Target condition being diagnosed

Lumbar vertebral fractures can occur in younger people exposed to trauma (vertebral traumatic fracture) or overuse (vertebral stress fracture) and in older people with osteoporosis (osteoporotic vertebral compression fracture). These different types of vertebral fractures may have different distinguishing clinical histories and clinical features, and may require different diagnostic work-up and management.

Lumbar vertebral stress fractures in younger people are thought to be secondary to chronic low-grade trauma or repetitive loading, which generally occurs in sport (Alqarni 2015). Lumbar vertebral traumatic fracture generally results from high-energy trauma (Lee 2016; Marongiu 2018). Osteoporotic vertebral compression fractures on the other hand can occur from minor trauma or without any definitive trauma in osteoporotic older people (Lee 2016). Osteoporotic vertebral fractures are associated with a 2- to 8-fold risk of mortality (Schousboe 2016), and a 5-fold increased risk of sustaining a further vertebral fracture (Marongiu 2018), and future hip fracture (Warriner 2011). A systematic review of 14 primary care studies (n = 14,860) found the median prevalence of vertebral fractures in patients presenting with low back pain was estimated at 3.6% in primary care and 6.5% in secondary and tertiary care (Downie 2013).

Index test(s)

The purpose of the index test is screening for potential lumbar vertebral fracture. In patients presenting with low back pain, clinical practice guidelines generally recommend assessing for the following red flags to help identify patients with suspicion of vertebral fracture: major/significant trauma; use of steroids; > 50 years of age; and obvious structural deformity (Verhagen 2016).

Despite the inclusion of several red flags in guidelines, the usefulness of red flags to screen for serious pathology, such as fracture, continues to be debated (Grunau 2017; Underwood 2009). Several red flags endorsed by previous guidelines have poor or untested diagnostic accuracy with little information on their utility for clinical practice (Cook 2018), and are based on evidence from individual studies (Parreira 2019).

In 2013, our group published a Cochrane Review of eight studies which had evaluated a total of 29 different types of index tests used to screen patients presenting with low back pain for vertebral fracture (Williams 2013). Results from individual studies suggest many of the endorsed red flags, such as thoracic pain and neurological deficit, were uninformative because these red flags had a positive and negative likelihood ratio very close to 1 (Williams 2013). There was also variation between studies in estimates of the diagnostic accuracy of index tests. For example, for significant trauma, the sensitivity ranged from 0.25 (95% confidence interval (CI) 0.03 to 0.65) to 0.65 (95% CI 0.44 to 0.83) and the specificity ranged from 0.90 (95% CI 0.86 to 0.93) to 0.98 (95% CI 0.96 to 0.98). For corticosteroid use the sensitivity ranged from 0.00 (95% CI 0.00 to 0.23) to 0.25 (95% CI 0.03 to 0.65), however the test was highly specific (0.99, 95% CI 0.98 to 1.00). For older age (> 74 years), the sensitivity ranged from 0.25 to 0.59 and the specificity ranged from 0.84 to 0.97. The positive likelihood ratios for the different red flags were: significant trauma - from 3.42 (95% CI 1.57 to 7.45) to 12.85 (95% CI 8.58 to 19.24); older age - from 3.69 (95% CI 3.00 to 4.53) to 9.39 (95% CI 2.69 to 32.75); and corticosteroid use - from 3.97 (95% CI 0.20 to 79.15) to 48.50 (95% CI 11.46 to 204.98).

Red flags used in isolation had modest diagnostic accuracy, considering they are required to detect a condition that has a low prevalence. Combinations of red flags seemed more informative than red flags used alone. For example age > 64 had a positive likelihood ratio of 7.13 which increased to 14.59 when combined with female gender (Williams 2013). One red flag in tertiary care appeared informative with the positive likelihood ratio high for contusion/abrasion (31.09, 95% CI 18.25 to 52.96). There was uncertainty about the value of most red flags, as estimates were imprecise, there was a lack of operational definition for older age cut-off points between studies, and most red flags were only evaluated by one study.

Clinical pathway

Most clinical practice guidelines recommend history and physical examination to identify patients with a higher likelihood of low back pain due to vertebral fracture (Oliveira 2018), who require further diagnostic work up. Different types of vertebral fractures may require different imaging procedures.

The American College of Radiography (ACR) recommends standard radiography for initial imaging in patients presenting with low back pain with a suspicion of osteoporotic vertebral compression

fracture (Patel 2016). Genant's semi-quantitative assessment and the algorithm-based qualitative approach are methods used to classify and confirm the presence of a vertebral fracture on standard radiographs (Genant 1993; Jiang 2004). Computed tomography (CT) is recommended over a plain radiograph if there is suspicion of vertebral traumatic fracture or vertebral stress fracture (Patel 2016).

Due to concerns of radiation exposure from standard radiography and CT (Berrington de González 2009; Wall 2006), vertebral fracture assessment (VFA) by dual-energy x-ray absorptiometry (DXA) is a potential alternative to assess osteoporotic and traumatic vertebral fractures (Lee 2016). However, there is currently insufficient evidence to replace standard radiography with VFA due to its reduced ability to detect lower-grade vertebral fractures (Fuerst 2009; Lee 2016; Lems 2021), reduced image resolution (Bazzocchi 2012; Damiano 2006; Schousboe 2006), and numerous potential sources of error (Johansson 2020).

Magnetic resonance imaging (MRI) is now preferred over CT in evaluating patients with suspicion of lumbar stress fracture. MRI can detect bone marrow oedema early, facilitating early diagnosis of lumbar stress reaction and fracture without radiation exposure (Ang 2016; Astur 2015). Given that, most patients presenting with low back pain with suspicion of lumbar stress fracture tend to be from younger age groups, MRI may be beneficial as first-line imaging given its high sensitivity and specificity, and no ionising radiation exposure (Alqarni 2015; Astur 2015; Cheung 2018; Murthy 2012). However, the high cost of MRI and potential increase in downstream expenditure need to be considered (Jarvik 2003).

Alternative test(s)

Apart from a combination of history taking and physical examination, there are no alternative screening tests for vertebral fracture in patients presenting with low back pain. Routinely imaging all patients is possible but is discouraged in current guidelines (Oliveira 2018).

Rationale

In light of the recent publication of primary diagnostic studies (de Schepper 2016; Premkumar 2018), and due to the evolving guidance for the most appropriate methods to systematically review studies of diagnostic test accuracy (Deeks 2022; Whiting 2011), we aim to update the previous Cochrane Review (Williams 2013). The results of this review will provide researchers and clinicians with more clarity in the use of red flags when screening for vertebral fracture in patients presenting with low back pain.

OBJECTIVES

The objective of this review is to assess the diagnostic accuracy of red flags used to screen for vertebral fracture in patients presenting with low back pain. Where possible, we will report the results of red flags separately for different types of vertebral fracture as follows.

- Osteoporotic vertebral compression fracture
- Vertebral traumatic fracture
- Vertebral stress fracture

Secondary objectives

The secondary objective of this review is to assess the influence of sources of heterogeneity on the diagnostic accuracy of red flag tests, if appropriate. In particular, we will assess the healthcare setting (e.g. primary, emergency, or secondary care) and the different types of vertebral fracture (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, and vertebral stress fracture. We will also assess mean age at different cut-off thresholds (e.g. > 50 years of age versus > 70 years of age) across different studies, as this has previously been shown to influence results.

METHODS

Criteria for considering studies for this review

Types of studies

We will consider primary diagnostic studies if they compare the results of history taking and/or physical examination findings (index test) with a reference standard test for the identification of vertebral fracture in patients presenting with low back pain. We will include studies that report results of red flags separately for the different types of vertebral fracture or studies that report combined results for vertebral fractures. We will include studies including cross-sectional designs that present sufficient data to allow estimates of diagnostic accuracy (such as sensitivity and specificity) to be derived. We will contact authors of abstracts and conference proceedings for the full text; if the full text is unavailable, we will exclude such studies.

We will exclude case-control studies (i.e. two-gate studies) that have included some participants (cases) already diagnosed with vertebral fracture and some participants (controls) without vertebral fracture as these types of studies may overestimate both sensitivity and specificity (Whiting 2013).

We will retrieve and include studies published in all languages. If necessary, we will organise appropriate translation of potentially eligible articles.

Participants

We will include studies if they evaluate patients presenting to primary, secondary or tertiary care for management of low back pain. We will exclude paediatric populations as presentation, risk factors, and fracture patterns will likely differ compared to adult populations.

Index tests

Studies evaluating any aspects of the history and/or physical examination for patients presenting with low back pain will be eligible for inclusion. This includes all information regarding:

- demographic characteristics (e.g. age, gender);
- clinical and medical history (e.g. pain intensity, previous history of falls); and
- physical examination results (e.g. tenderness on palpation, muscle strength testing, or lumbar range of motion).

We will include studies if the diagnostic accuracy of individual red flags are evaluated in isolation, or as part of a combination (e.g. female sex, > 70 years of age, significant trauma, and prolonged

steroid use). We will extract or derive indices of diagnostic performance from data presented in each primary study for each red flag or combination of red flags. We will exclude studies where only a 'clinical diagnosis' or 'global clinician judgement' were used (without specifying which diagnostic tools were used) as clinical judgement based on undefined parameters is difficult to interpret.

Target conditions

We will include all studies that investigated patients presenting with low back pain or for lumbar examination and report appropriate data to calculate an estimate of the diagnotic accuracy of the prevalence of vertebral fracture. If possible, we will describe results separately for different types of vertebral fracture (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, and vertebral stress fracture).

Where possible, we will attempt to categorise the different types of vertebral fractures under the following definitions.

- Osteoporotic vertebral compression fracture: insufficiency fractures (i.e. a fracture due to normal stresses on an abnormal bone).
- Vertebral traumatic fracture: any vertebral fracture due to a trauma mechanism in normal bone (e.g. direct force to the lumbar spine).
- Vertebral stress fracture: fractures as a result of repetitive trauma over time (i.e. a fracture due to abnormal stresses on a normal bone).

Reference standards

Where possible, we will extract the reference standard used to confirm the different types of vertebral fracture separately. If studies report sufficient data, we will explore how the different reference standards impact the diagnostic accuracy of the index test. For traumatic vertebral fractures, CT will be the preferential reference test. If CT scan is not available, we will use MRI as the reference test. If neither are used, we will use whichever imaging modality is presented as the reference standard. For osteoporotic vertebral fractures and vertebral stress fractures, MRI will be the preferential reference test. If MRI is not available, we will use bone scintigraphy with SPECT as the reference test. If neither are used, we will use whichever imaging modality is presented as the reference standard. However, we anticipate that most studies will not separate reference standards based on the different types of vertebral fracture.

We will also consider clinical follow-up after the initial consultation as a weaker reference standard if suspected vertebral fracture was subsequently confirmed by imaging.

Search methods for identification of studies

Electronic searches

The search strategy will be based upon the Williams 2013 Cochrane Review and updated in collaboration with a librarian. We will search relevant databases for eligible diagnostic studies from the Williams 2013 search date onwards. We will search the following databases.

- MEDLINE (Ovid, from 1946) (Appendix 1)
- Embase (Appendix 2)
- CINAHL (Appendix 3)

Searching other resources

We will check the reference lists of all included publications. Forward citation tracking of the included studies will also be performed through Web of Science, Scopus, and Google Scholar. We will send the final list of included studies for review by experts in the field to minimise the risk of missing any relevant studies.

Data collection and analysis

Selection of studies

Two review authors (CSH, AD) will independently apply the selection criteria to all studies (titles and abstracts) identified from the literature search. We will exclude clearly irrelevant studies. Two review authors (CSH, AD) will retrieve and independently assess full-text publications using the predefined inclusion criteria. All disagreements will be resolved by consensus, and other review authors will be consulted if necessary. We will contact relevant study authors via email if any study information needs to be clarified. We will record all excluded studies with reasons for their exclusion and illustrate the study selection process in a PRISMA diagram (McInnes 2018).

Data extraction and management

Two review authors (CSH, AD) will independently extract the relevant data from the included studies. This will include data on the characteristics of the included studies such as:

- participant characteristics;
- type of vertebral fracture, if reported;
- types of index tests and reference standards and their descriptors;
- relevant aspects of the study methods;
- data for the assessment of diagnostic accuracy.

Characteristics of participants (and studies) will include details on:

- healthcare setting, e.g. primary care (medical centre, physiotherapist, chiropractor), emergency departments, secondary care (outpatient hospital service, medical hospital services, specialist services);
- inclusion and exclusion criteria;
- enrolment procedures (consecutive or non-consecutive);
- number of patients (including number eligible and enrolled in the study);
- number of patients receiving the index test and reference standard;
- number of patients for whom results are reported in the two-by-two table;
- reasons for withdrawal;
- patient demographics (age, gender, duration, and history of low back pain).

Test characteristics will include:

- type of index test;
- methods of execution of index tests;
- threshold for positive result on an index test;
- experience and expertise of the assessors;
- type of reference standard;

- cut-off points for diagnosing vertebral fracture on a reference standard (e.g. quantitative radiographic measures) where relevant.

Thresholds for 'positive' results on an index test may vary across studies and some studies may present the diagnostic performance of an index test at several different cut-off points (e.g. mean age > 50 years versus mean age > 70 years). This will be displayed on a Receiver Operating Characteristic (ROC) curve. We will extract all cut-off points from the included studies.

To assess diagnostic accuracy, we will extract diagnostic two-by-two tables (true positive, false positive, true negative, and false negative counts) from the publications or reconstruct them using information from other relevant parameters (sensitivity, specificity, or predictive values). If the reported sensitivity, specificity, positive and negative predictive values do not match the reported two-by-two table data we will report results based on the raw two-by-two data. However, if there are large discrepancies in the values calculated from raw two-by-two data and the values presented in a study, we will first contact the study authors to seek clarification of the data. If extracting raw data to produce a two-by-two table is not possible, we will then use the sensitivity and specificity values reported. Any discrepancies in calculated two-by two data and values presented in a study will be discussed with the author team and will be reported in the manuscript. For eligible studies where we cannot reconstruct the diagnostic two-by-two tables, we will present them in the review but will not include them in the quantitative analysis. Two review authors (CSH, AD) will independently extract the data to ensure adequate reliability of collected data. For each study, if possible, we will present in tables aspects of study design, characteristics of the population, index test, and reference standard, and type of vertebral fracture. Where a review author is also an author of one of the primary diagnostic studies, they will not be involved in the data extraction or quality rating of that study.

Assessment of methodological quality

Two review authors (CSH, AD) will independently assess the methodological quality of each study using the QUADAS-2 checklist (Whiting 2011), as recommended by the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Appendix 4; Deeks 2022). The QUADAS-2 checklist consists of four domains: patient selection, index test, reference standard, and flow and timing. We will assess all domains for risk of bias and the three domains for concerns regarding applicability.

The same review authors will classify each item as 'yes' (adequately addressed); 'no' (inadequately addressed); or 'unclear' (inadequate detail presented to allow a judgement to be made). Disagreements will be resolved by discussion and if necessary, by consulting a third review author (CGM or MJH).

Statistical analysis and data synthesis

Statistical analysis will follow the approaches outlined in Chapter 10 of the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Deeks 2022). Indices of diagnostic performance will be extracted or derived from data presented in each primary study for each red flag or combination of red flags. Sensitivity and specificity of the index test will be the primary measures of diagnostic test accuracy, both in text and summary of findings tables. The summary of findings table will include the following variables, but not limited to: the review question, patient population, index test, target condition, reference standard, study quality, included studies characteristics, sensitivity, specificity, and likelihood ratios. We will generate diagnostic two-by-two tables, from which we will calculate and present sensitivities and specificities for each index test with 95% CIs and present these in forest plots. We will calculate likelihood ratios for each test and use these as an indication of clinical usefulness. In addition, we will use a ROC plot of sensitivity versus 1-specificity to display the data. We will present results of how different pre-test probability cut-offs and positive or negative results impact post-test probabilities.

We will use the bivariate random-effects model to compute summary estimates of sensitivity and specificity. This method accounts for both within-study and between-study variability and allows for correlation that may exist between sensitivity and specificity across studies (Harbord 2007; Reitsma 2005). We will assume a binomial distribution within studies (Chu 2006), and we will assume the random effects to be normally distributed across studies. We will restrict these analyses to studies that show sufficient clinical homogeneity (e.g. same index test and index test threshold, and a similar definition of vertebral fracture). We will present summary estimates with a 95% confidence ellipse (i.e. a bivariate confidence interval) in ROC space. We will use pooled estimates of sensitivity and specificity to calculate the positive and negative likelihood ratios for each index test. We will use SAS version 9.4 (SAS Institute 2013) and Review Manager 5 software (Review Manager 2020).

For index tests where the cut-off for test positivity varies across studies, we will use the Hierarchical Summary ROC (HSROC) method of Rutter and Gatsonis (Rutter 2001) to estimate a summary curve which we will display graphically. Where a study has data for more than one cut-off point, we will use the most commonly reported cut-off point across studies. We will conduct the HSROC analyses in SAS (SAS Institute 2013).

Investigations of heterogeneity

Several factors can contribute to the heterogeneity in the diagnostic accuracy of a test across multiple studies. We will investigate the potential influences of one study level variable at a time, such as healthcare setting (i.e. primary care, accident and emergency, secondary care) and different types of vertebral fracture (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, and vertebral stress fracture). We chose the healthcare setting as the prevalence of vertebral fracture may vary depending on healthcare setting. We chose different types of vertebral fracture as the diagnostic accuracy of red flags may vary dependent on the type of fracture. We will aim to pool results separately for different healthcare settings and for each of the different types of vertebral fracture, if data are available. If results of independent red flags are not reported separately, we will report results as a combination. Where feasible, we will include study level covariates (e.g. different types of imaging) in the HSROC analysis models to investigate sources of heterogeneity. Where appropriate we will clearly acknowledge any unexplained heterogeneity between studies in the summary of findings table, as per recommendations of the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Deeks 2022).

Sensitivity analyses

We will perform a sensitivity analysis to investigate how individual QUADAS-2 key domains (i.e. patient selection, index test, reference standard, and flow and timing) affect accuracy estimates. We will assess whether estimates of sensitivity and specificity are affected if studies at high risk of bias in either of the QUADAS-2 domains are removed for the sensitivity analysis. If possible, we will also assess whether estimates of sensitivity and specificity for diagnostic accuracy differ for the different types of vertebral fracture.

Assessment of reporting bias

Tests for publication bias in reviews of diagnostic test accuracy are not routinely used because study size may be related to test accuracy for reasons other than publication bias (Deeks 2005). Therefore, we will not report on this.

ACKNOWLEDGEMENTS

We would like to thank Petra Macaskill and Maurits van Tulder for their assistance in the development of this protocol. We acknowledge peer reviewers: Prof Roger Chou, MD, Oregon Health & Science University.

REFERENCES

Additional references

Alqarni 2015

Alqarni AM, Schneiders AG, Cook CE, Hendrick PA. Clinical tests to diagnose lumbar spondylolysis and spondylolisthesis: A systematic review. *Physical Therapy in Sport* 2015;**16**(3):268-75.

Ang 2016

Ang EC, Robertson AF, Malara FA, O'Shea T, Roebert JK, Schneider ME, et al. Diagnostic accuracy of 3-T magnetic resonance imaging with 3D T1 VIBE versus computer tomography in pars stress fracture of the lumbar spine. *Skeletal Radiology* 2016;**45**(11):1533-40.

Astur 2015

Astur DC, Zanatta F, Arliani GG, Moraes ER, Pochini Ade C, Ejnisman B. Stress fractures: definition, diagnosis and treatment. *Revista Brasileira de Ortopedia* 2015;**51**(1):3-10.

Bazzocchi 2012

Bazzocchi A, Spinnato P, Fuzzi F, Diano D, Morselli-Labate AM, Sassi C, et al. Vertebral fracture assessment by new dual-energy X-ray absorptiometry. *Bone* 2012;**50**(4):836-41.

Berrington de González 2009

Berrington de González A, Mahesh M, Kim KP, Bhargavan M, Lewis R, Mettler F, et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. *Archives of Internal Medicine* 2009;**169**(22):2071-7.

Buchbinder 2018

Buchbinder R, van Tulder M, Öberg B, Costa LM, Woolf A, Schoene M, et al, Lancet Low Back Pain Series Working Group. Low back pain: a call for action. *Lancet* 2018;**391**(10137):2384-8.

Cheung 2018

Cheung KK, Dhawan RT, Wilson LF, Peirce NS, Rajeswaran G. Pars interarticularis injury in elite athletes - The role of imaging in diagnosis and management. *European Journal of Radiology* 2018;**108**:28-42.

Chu 2006

Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006;**59**(12):1331-3.

Cook 2018

Cook CE, George SZ, Reiman MP. Red flag screening for low back pain: nothing to see here, move along: a narrative review. *British Journal of Sports Medicine* 2018;**52**(8):493-6.

Damiano 2006

Damiano J, Kolta S, Porcher R, Tournoux C, Dougados M, Roux C. Diagnosis of vertebral fractures by vertebral fracture assessment. *Journal of Clinical Densitometry* 2006;**9**(1):66-71.

Deeks 2005

Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* 2005;**8**(9):882-93.

Deeks 2022

Deeks JJ, Bossuyt PM, Leeflang M, Takwoingi Y, Flemyng E. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.0. The Cochrane Collaboration, 2022. Available from methods.cochrane.org/sdt/handbook-dta-reviews.

DePalma 2020

DePalma MG. Red flags of low back pain. *Journal of the American Academy of PAs* 2020;**33**(8):8-11. [DOI: 10.1097/01.JAA.0000684112.91641.4c]

de Schepper 2016

de Schepper EI, Koes BW, Veldhuizen EF, Oei EH, Bierma-Zeinstra SM, Luijsterburg PA. Prevalence of spinal pathology in patients presenting for lumbar MRI as referred from general practice. *Family Practice* 2016;**33**(1):51-6.

Downie 2013

Downie A, Williams CM, Henschke N, Hancock MJ, Ostelo RW, de Vet HC, et al. Red flags to screen for malignancy and fracture in patients with low back pain: systematic review. *BMJ* 2014;**48**(20):1518-56.

Ferreira 2021

Ferreira GE, Herbert R, Machado GC, Richards B, Harris IA, O'Keeffe M, et al. Low back pain presentations to New South Wales emergency departments: Trends over time and geographical variation. *Emergency Medicine Australasia* 2021;**33**(5):868-74.

Fuerst 2009

Fuerst T, Wu C, Genant HK, von Ingersleben G, Chen Y, Johnston C, et al. Evaluation of vertebral fracture assessment by dual X-ray absorptiometry in a multicenter setting. *Osteoporosis International* 2009;**20**(7):1199–205.

Galliker 2019

Galliker G, Scherer DE, Trippolini MA, Rasmussen-Barr E, LoMartire R, Wertli MM. Low back pain in the emergency department: prevalence of serious spinal pathologies and diagnostic accuracy of red flags. *American Journal of Medicine* 2019;**133**(1):60-72.

Genant 1993

Genant HK, Wu CY, van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. *Journal of Bone and Mineral Research* 1993;**8**(9):1137-48.

Grunau 2017

Grunau GL, Darlow B, Flynn T, O'Sullivan K, O'Sullivan PB, Forster BB. Red flags or red herrings? Redefining the role of red flags in low back pain to reduce overimaging. *British Journal of Sports Medicine* 2017;**52**(8):488-9.

Harbord 2007

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

Hartvigsen 2018

Hartvigsen J, Hanock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, et al. What low back pain is and why we need to pay attention. *Lancet* 2018;**391**(10137):2356-67.

Henschke 2009

Henschke N, Maher CG, Refshauge KM, Herbert RD, Cumming RG, Bleasel J, et al. Prevalence of and screening for serious spinal pathology in patients presenting to primary care settings with acute low back pain. *Arthritis & Rheumatology* 2009;**60**(10):3072-80.

Jarvik 2003

Jarvik JG, Hollingworth W, Martin B, Emerson SS, Gray DT, Overman S, et al. Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomised controlled trial. *JAMA* 2003;**289**(21):2810–18.

Jiang 2004

Jiang, G, Eastell R, Barrington NA, Ferrar L. Comparison of methods for the visual identification of prevalent vertebral fracture in osteoporosis. *Osteoporosis International* 2004;**15**(11):887–96.

Johansson 2020

Johansson L, Sundh D, Magnusson P, Rukmangatharajan K, Mellström D, Nilsson AG, et al. Grade 1 vertebral fractures identified by densitometric lateral spine imaging predict incident major osteoporotic fracture independently of clinical risk factors and bone mineral density in older women. *Journal of Bone and Mineral Research* 2020;**35**(10):1942-51.

Koes 2006

Koes BW, van Tulder MW, Thomas S. Diagnosis and treatment of low back pain. *BMJ* 2006;**332**(7555):1430-4.

Lee 2016

Lee JH, Lee YK, Oh SH, Ahn J, Lee E, Pyo JH, et al. A systematic review of diagnostic accuracy of vertebral fracture assessment (VFA) in postmenopausal women and elderly men. *Osteoporosis International* 2016;**27**(5):1691-9.

Lems 2021

Lems WF, Paccou J, Zhang J, Fuggle NR, Chandran M, Harvey NC, et al. Vertebral fracture: epidemiology, impact and use of DXA vertebral fracture assessment in fracture liaison services. *Osteoporosis International* 2021;**32**(3):399–411.

Marongiu 2018

Marongiu G, Congia S, Verona M, Lombardo M, Podda D, Capone A. The impact of magnetic resonance imaging in the diagnostic and classification process of osteoporotic vertebral fractures. *Injury* 2018;**49**(Suppl 3):S26-S31.

McInnes 2018

McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, PRISMA-DTA Group. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* 2018;**319**(4):388-96.

Murthy 2012

Murthy NS. Imaging of stress fractures of the spine. *Radiologic Clinics of North America* 2012;**50**(4):799-821.

Oliveira 2018

Oliveira CB, Maher CG, Pinto RZ, Traeger AC, Lin CW, Chenot JF, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *European Spine Journal* 2018;**27**:2791–803.

Parreira 2019

Parreira PC, Maher CG, Traeger AC, Hancock MJ, Downie A, Koes BW, et al. Evaluation of guideline-endorsed red flags to screen for fracture in patients presenting with low back pain. *British Journal of Sports Medicine* 2019;**53**(10):648-54.

Patel 2016

Patel ND, Broderick DF, Burns J, Deshmukh TK, Blair IB, Harvey B, et al. ACR Appropriateness Criteria Low Back Pain. *Journal of the American College of Radiology* 2016;**13**(9):1069-78.

Premkumar 2018

Premkumar A, Godfrey W, Gottschalk MB, Boden SD. Red flags for low back pain are not always really red: a prospective evaluation of the clinical utility of commonly used screening questions for low back pain. *Journal of Bone and Joint Surgery* 2018;**100**(5):368-74.

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.

Review Manager 2020 [Computer program]

Nordic Cochrane Centre, The Cochrane Collaboration Review Manager 5 (RevMan 5). Version 5.4. Copenhagen: Nordic Cochrane Centre, The Cochrane Collaboration, 2020.

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.

SAS Institute 2013 [Computer program]

SAS Institute SAS Version 9.4. SAS Institute. Cary, North Carolina: SAS Institute, 2013.

Schousboe 2006

Schousboe JT, Ensrud KE, Nyman JA, Kane RL, Melton LJ. Cost-effectiveness of vertebral fracture assessment to detect prevalent vertebral deformity and select postmenopausal women with a femoral neck T-score>-2.5 for alendronate

therapy: a modelling study. *Journal of Clinical Densitometry* 2006;**9**(2):133-43.

Schousboe 2016

Schousboe JT. Epidemiology of vertebral fractures. *Journal of Clinical Densitometry* 2016;**19**(1):8-22.

Underwood 2009

Underwood M. Diagnosing acute nonspecific low back pain: time to lower the red flags? *Arthritis & Rheumatology* 2009;**60**(10):2855-7.

Verhagen 2016

Verhagen AP, Downie A, Popal N, Maher C, Koes BW. Red flags presented in current low back pain guidelines: a review. *European Spine Journal* 2016;**25**(9):2788-802.

Wall 2006

Wall BF, Kendall GM, Edwards AA, Bouffler S, Muirhead CR, Meara JR. What are the risks from medical X-rays and other low dose radiation? *British Journal of Radiology* 2006;**79**(940):285-94.

Warriner 2011

Warriner AH, Patkar NM, Curtis JR, Delzell E, Gary L, Kilgore M, et al. Which fractures are most attributable to osteoporosis? *Journal of Clinical Epidemiology* 2011;**64**(1):46-53.

Whiting 2011

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011;**155**(8):529-36.

Whiting 2013

Whiting PF, Rutjes AW, Westwood ME, Mallett S, QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology* 2013;**66**(10):1093-104.

Williams 2013

Williams CM, Henschke N, Maher CG, van Tulder MW, Koes BW, Macaskill P, et al. Red flags to screen for vertebral fracture in patients presenting with low-back pain. *Cochrane Database of Systematic Reviews* 2013, Issue 1. Art. No: CD008643. [DOI: 10.1002/14651858.CD008643.pub2]

Wu 2020

Wu A, March L, Zheng X, Huang J, Wang X, Zhao J, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. *Annals of Translational Medicine* 2020;**8**(6):299. [DOI: 10.21037/atm.2020.02.175.]

APPENDICES

Appendix 1. MEDLINE search strategy

I. Index test: clinical red flags

1. exp Medical History Taking/

2. exp Physical Examination/

3. exp "Wounds and Injuries"/

4. exp Accidental Falls/

5. exp Accidents/

6. physical examination*.tw,kf

7. history.tw,kf

8. red flag*.tw,kf

9. physical test*.tw,kf

10. trauma*.tw,kf

11. injur*.tw,kf

12. (clinical* adj (diagnosis or sign* or significance or symptom* or parameter* or assessment* or finding* or evaluat* or indication* or examination*)).tw,kf

13. OR/1–12

II. Population: low back pain

14. exp Back Pain/

15. exp Low Back Pain/

16. exp Sciatica/

17. exp Sciatic Neuropathy/

18. dorsalgia*.tw,kf

19. backache*.tw,kf

20. lumbago.tw,kf

21. back disorder*.tw,kf

22. (lumb* adj pain).tw,kf

23. coccyx.tw,kf

24. coccydynia.tw,kf

25. sciatica.tw,kf

26. spondylosis.tw,kf

27. OR/14–26

III. Target condition: vertebral fracture

28. Fractures, Bone/

29. Fractures, Stress/

30. Fractures, Spontaneous/

31. Fractures, Compression/

32. Fractures, Closed/

33. Spinal Injuries/

34. Spinal Diseases/

35. fracture*.tw,kf

36. OR/28–35

IV. Limits and search combination

37. (20120401-20170517).ed

38. (2012-2017).yr

39. OR/38–39

40. 13 AND 27 AND 36 AND 40

Appendix 2. Embase search strategy

1. exp Anamnesis/

2. exp Physical Examination/

3. exp Physical Performance/

4. exp Accident/

5. physical examination*.tw,kw

6. history.tw,kw

7. red flag*.tw,kw

8. physical test*.tw,kw

9. trauma*.tw,kw

10. injur*.tw,kw

11. (clinical* adj (diagnosis or sign* or significance or symptom* or parameter* or assessment* or finding* or evaluat* or indication* or examination*)).tw,kw

12. OR/1–11

II. Population: low back pain

13. exp Low back Pain/

14. exp Backache/

15. exp Spondylosis/

16. exp Sciatica/

17. exp Ischialgia/

18. dorsalgia.tw,kw

19. back pain.tw,kw

20. backache.tw,kw

21. lumbago.tw,kw

22. back disorder*.tw,kw

23. (lumb* adj pain).tw,kw

24. spondylosis.tw,kw

25. coccyx.tw,kw

26. coccydynia.tw,kw

27. sciatica.tw,kw

28. OR/13–27

III. Target condition: vertebral fracture

29. exp Fracture/

30. exp Spine Fracture/

31. exp Bone Injury/

32. exp Spine Injury/

33. exp Spine Disease/

34. fracture*.tw,kw

35. OR/29–34

IV. Limits and search combination

36. (20120401-20170517).dd

37. (2012-2017).yr

38. OR/36–37

39. 12 AND 28 AND 35 AND 38

40. limit 39 to embase

Appendix 3. CINAHL search strategy

I. Index test: clinical red flags

1. (MH "Patient History Taking")

2. (MH "Diagnostic Tests, Routine")

3. (MM "Physical Examination")

4. (TX "history" or "red flag*" or "physical examination" or "physical test" or "clinical*" or "diagnosis" or TX "sign" or TX "signs" or TX "significance" or TX "symptom*" or TX "parameter*" or TX "assessment" or TX "finding*" or TX "evaluat*" or TX "indication*" or TX "examination*")

5. OR/1–4

II. Population: low back pain

6. (MH "Back Pain+")

7. (MH "Low Back Pain")

8. (MH "Coccyx")

9. (MH "Sciatica")

10. (MH "Lumbar Vertebrae")

11. (MH "Spondylolisthesis")

12. (MH "Spondylolysis")

13. (TX "back pain")

14. (TX "backache")

15. (TX "dorsalgia")

16. (TX "lumbago")

17. (TX "sciatica")

18. (TX "coccyx")

19. (TX "coccydynia")

20. (TX "back disorder*")

21. (TX "lumbago")

22. lumb* w1 pain

23. lumb* n5 pain

24. lumb* n2 vertebra*

25. OR/11–29

III. Target condition: vertebral fracture

26. (MH "Fractures, Closed") or (MH "Fractures, Compression+") or (MH "Fractures, Spontaneous") or (MH "Fractures, Stress+")

27. (MH "Spinal Fractures+")

28. (MH "Spinal Injuries")

29. (MH "Spinal Diseases")

30. (MH "Wounds and Injuries")

31. (TX "fracture*")

32. (TX "trauma")

33. (TX "injury")

34. OR/26–33

IV. Limits and search combination

35. EM 201204-

36. DT 2012-

37. OR/35-36

38. 5 AND 25 AND 34 AND 37

Appendix 4. QUADAS-2 tool: Risk of bias and applicability judgements

DOMAIN 1: PATIENT SELECTION	
A. Risk of bias	
Describe methods of patient selection:	
• **Was a consecutive or random sample of patients enrolled?** **Score 'yes'** if a consecutive series of patients or a random sample has been selected. Details regarding setting, in inclusion and exclusion criteria, and preferably number of patients eligible and excluded should be given. If a mixed population (e.g. primary and secondary care) is used, the number of participants from each setting will be presented. **Score 'no'** if healthy controls are used. If non-response is high and selective, or there is clear evidence of selective sampling, this will be scored 'no'. **Score 'unclear'** if insufficient information is given on the setting, selection criteria, or selection procedure to make a judgement.	Yes/No/Unclear
• **Did the study avoid inappropriate exclusions?** **Score 'yes'** if participants were excluded based on the prespecified exclusion criteria or if the exclusion criteria was clearly appropriate for the study population. **Score 'no'** if inappropriate exclusion criteria are used such as excluding based on gender, or inappropriate selection bias occurs. **Score 'unclear'** if insufficient information is provided to assess this item.	Yes/No/Unclear
Could the selection of patients have introduced bias? **Score 'low'** if all questions are rated as 'yes'. **Score 'high'** if any of the questions is rated as 'no'.	Risk: Low/High/Unclear

(Continued)

Score 'unclear if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is sufficient information to make a decision about the risk of bias.

B. Concerns regarding applicability

Describe included patients (prior testing, presentation, intended use of index test and setting):

We expect that most of the diagnostic studies will include participations appropriate for the review question and would have used the index tests described in our inclusion criteria. Many of the index tests are applicable across the different healthcare settings.

Is there concern that the included patients do not match the review question? Concern: Low/High/Unclear

Score 'low concern' if participants meet all inclusion criteria or meet 3 out of the 4 criteria.

Score 'high concern' if participants meet 1 or less of the 4 criteria.

Score 'unclear' if insufficient information is provided to assess this item.

DOMAIN 2: INDEX TEST(S) (if more than 1 index test was used, please complete for each test)

A. Risk of bias

Describe the index test and how it was conducted and interpreted:

- **Were the index test results interpreted without knowledge of the results of the reference standard?** Yes/No/Unclear

Score 'yes' if the results of the index test are interpreted blind to the results of the reference test or if the sequence of testing is always the same (i.e. the index test is always performed first, followed by the reference standard), and consequently the index test is interpreted blind of the reference standard.

Score 'no' if the assessor is aware of the results of the reference standard.

Score 'unclear' if insufficient information is given on independent or blind assessment of the reference standard.

- **If a threshold was used, was it prespecified?** Yes/No/Unclear

Score 'yes' if it is clear that a prespecified threshold was used e.g. the authors follow a study protocol.

Score 'no' if a prespecified threshold was not used or different thresholds were used and not predefined.

Score 'unclear' if insufficient information is provided to assess this item.

Could the conduct or interpretation of the index test have introduced bias? Risk: Low/High/Unclear

Score 'low' if all questions are rated as 'yes'.

Score 'high' if any of the questions is rated as 'no'.

Score 'unclear if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is enough information to make a decision about the risk of bias.

B. Concerns regarding applicability

Is there concern that the index test, its conduct, or interpretation differ from the review question? Concern: Low/High/Unclear

(Continued)

Score 'low concerns' if the index text performed is the same as specified in the review question and was intended to assess vertebral fracture.

Score 'high concerns' if the index test performed does not follow what is specified in the review question and was not intended to assess vertebral fracture.

Score 'unclear' if insufficient information is provided to assess this item.

DOMAIN 3: REFERENCE STANDARD	
A. Risk of bias	
Describe the reference standard and how it was conducted and interpreted:	
• **Is the reference standard likely to correctly classify the target condition?** **Score 'yes'** if the following criteria is followed. For traumatic vertebral fractures, CT is preferentially used as the reference test. If CT scan is not available, MRI is used as the reference test. For osteoporotic vertebral fractures and vertebral stress fractures, MRI is preferential used as the reference test. If MRI is not available, bone scintigraphy with SPECT is used as the reference test. **Score 'no'** if consensus among observers, or an unknown combination of the clinical assessment ('clinical judgement') is used as reference standard. **Score as 'unclear'** if insufficient information is given on the reference standard to make an adequate assessment.	Yes/No/Unclear
• **Were the reference standard results interpreted without knowledge of the results of the index test?** **Score 'yes'** if the results of the reference standard are interpreted blind to the results of the index tests or if the sequence of testing is always the same (i.e. the reference standard is always performed first, followed by the index test) and consequently, the reference standard is interpreted blind of the index test. **Score 'no'** if the assessor is aware of the results of the index test. **Score 'unclear'** if insufficient information is given on independent or blind assessment of the index test.	Yes/No/Unclear
Could the reference standard, its conduct, or its interpretation have introduced bias? **Score 'low'** if all questions are rated as 'yes'. **Score 'high'** if any of the questions is rated as 'no' **Score 'unclear** if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is sufficient information to make a decision about the risk of bias.	Risk: Low/High/Unclear
B. Concerns regarding applicability	
Is there concern that the target condition as defined by the reference standard does not match the review question? **Score 'low' concern if the target condition defined by the reference standard matches the target condition specified in the review.** **Score 'high' concern if the target condition defined by the reference standard does not match the target condition specified in the review.** **Score 'unclear'** if insufficient information is provided to assess this item.	Concern: Low/High/Unclear

(Continued)

DOMAIN 4: FLOW AND TIMING	
A. Risk of bias	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):	
Describe the time interval and any interventions between index test(s) and reference standard:	
• **Was there an appropriate interval between index test(s) and reference standard?** **Score 'yes'** if the time period between clinical assessment and the reference standard ≤ 1 week. **Score 'no'** if the time period between clinical assessment and the reference standard is > 1 week. **Score 'unclear'** if there is insufficient information on the time period between index tests and reference standard.	Yes/No/Unclear
• **Did all patients receive the same reference standard?** **Score 'yes'** if it is clear that all patients receiving the index test are subjected to the same reference standard. **Score 'no'** if different reference standards are used or if only some patients are subjected to the reference standard. **Score 'unclear'** if insufficient information is provided to assess this item.	Yes/No/Unclear
• **Were all patients included in the analysis?** **Score 'yes'** if it is clear that all recruited patients (> 95%) were included in the analysis. **Score 'no'** if the number of patients recruited differs (> 5% difference) from the number of patients included for analysis. **Score 'unclear'** if insufficient information is provided to assess this item.	Yes/No/Unclear
Could the patient flow have introduced bias? **Score 'low'** if all questions are rated as 'yes'. **Score 'high'** if any of the questions is rated as 'no'. **Score 'unclear** if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is sufficient information to make a decision about the risk of bias.	Risk: Low/High/Unclear

CT: computed tomography; MRI: magnetic resonance imaging; SPECT: single-photon emission computed tomography.

CONTRIBUTIONS OF AUTHORS

All authors contributed to discussions regarding the conception and design of the study. GCM wrote the first draft of the protocol, with help from the other authors. All authors read and approved the final manuscript.

DECLARATIONS OF INTEREST

CSH: none known

MJH: none known

AD: none known

JGJ declares the following: Association of University Radiologists (Fiduciary Officer), Wolters Klewer Health, Inc. (Independent Contractor - Consultant), springer publishing (Independent Contractor - Other), GE Healthcare (travel)

BWK: none known

GCM: none known

APV: none known

CMW: none known

CGM is a senior editor for Cochrane Musculoskeletal. Editors are required to conduct at least one Cochrane Review, which ensures that they are aware of the processes and commitment needed to conduct reviews.

SOURCES OF SUPPORT

Internal sources

- School of Public Health, The University of Sydney, Australia

External sources

- National Health and Medical Research Council, Australia
- Sydney Health Partners, Australia

Chapter Four:

Red flags to screen for vertebral fracture in patients presenting with low back pain (Review)

Statement from co-authors confirming authorship contribution of the PhD candidate

The co-authors of the paper *"Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Chen Q, Maher CG. Red flags to screen for vertebral fracture in people presenting with low back pain. Cochrane Database Syst Rev. 2023 Aug 24;8(8):CD014461. doi: 10.1002/14651858.CD014461.pub2. PMID: 37615643; PMCID: PMC10448864."* confirm that Christopher S Han has provided the following contributions to the study:

- Conception and design of the research
- Data acquisition
- Data analysis and interpretation of findings
- Writing of the manuscript and critical appraisal of the content

Christopher S Han _____ Date: 22/11/2023

Professor Chris G Maher _____ Date: 22/11/2023

Professor Mark J Hancock _____ Date: 22/11/2023

Cochrane Library

Cochrane Database of Systematic Reviews

Red flags to screen for vertebral fracture in people presenting with low back pain (Review)

Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Chen Q, Maher CG

Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Chen Q, Maher CG.
Red flags to screen for vertebral fracture in people presenting with low back pain.
Cochrane Database of Systematic Reviews 2023, Issue 8. Art. No.: CD014461.
DOI: 10.1002/14651858.CD014461.pub2.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Diagnostic Test Accuracy Review]

Red flags to screen for vertebral fracture in people presenting with low back pain

Christopher S Han[1], Mark J Hancock[2], Aron Downie[1,3], Jeffrey G Jarvik[4], Bart W Koes[5,6], Gustavo C Machado[1], Arianne P Verhagen[7], Christopher M Williams[8], Qiuzhe Chen[1], Christopher G Maher[1]

[1]Institute for Musculoskeletal Health, The University of Sydney and Sydney Local Health District, Sydney, Australia. [2]Department of Health Sciences, Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia. [3]Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia. [4]Departments of Radiology and Neurological Surgery, and the UW Clinical Learning, Evidence And Research (CLEAR) Center for Musculoskeletal Disorders, University of Washington School of Medicine, Seattle, USA. [5]Center for Muscle and Joint Health, University of Southern Denmark, Odense, Denmark. [6]Department of General Practice, Erasmus University Medical Center, Rotterdam, Netherlands. [7]Discipline of Physiotherapy, Graduate School of Health, University of Technology Sydney (UTS), Sydney, Australia. [8]School of Medicine and Public Health, University of Newcastle, Callaghan, Australia

Contact: Christopher S Han, christopher.han@sydney.edu.au.

Editorial group: Cochrane Back and Neck Group.
Publication status and date: New, published in Issue 8, 2023.

Citation: Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, Verhagen AP, Williams CM, Chen Q, Maher CG. Red flags to screen for vertebral fracture in people presenting with low back pain. *Cochrane Database of Systematic Reviews* 2023, Issue 8. Art. No.: CD014461. DOI: 10.1002/14651858.CD014461.pub2.

ABSTRACT

Background

Low back pain is a common presentation across different healthcare settings. Clinicians need to confidently be able to screen and identify people presenting with low back pain with a high suspicion of serious or specific pathology (e.g. vertebral fracture). Patients identified with an increased likelihood of having a serious pathology will likely require additional investigations and specific treatment. Guidelines recommend a thorough history and clinical assessment to screen for serious pathology as a cause of low back pain. However, the diagnostic accuracy of recommended red flags (e.g. older age, trauma, corticosteroid use) remains unclear, particularly those used to screen for vertebral fracture.

Objectives

To assess the diagnostic accuracy of red flags used to screen for vertebral fracture in people presenting with low back pain. Where possible, we reported results of red flags separately for different types of vertebral fracture (i.e. acute osteoporotic vertebral compression fracture, vertebral traumatic fracture, vertebral stress fracture, unspecified vertebral fracture).

Search methods

We used standard, extensive Cochrane search methods. The latest search date was 26 July 2022.

Selection criteria

We considered primary diagnostic studies if they compared results of history taking or physical examination (or both) findings (index test) with a reference standard test (e.g. X-ray, magnetic resonance imaging (MRI), computed tomography (CT), single-photon emission computerised tomography (SPECT)) for the identification of vertebral fracture in people presenting with low back pain. We included index tests that were presented individually or as part of a combination of tests.

Data collection and analysis

Two review authors independently extracted data for diagnostic two-by-two tables from the publications or reconstructed them using information from relevant parameters to calculate sensitivity, specificity, and positive (+LR) and negative (−LR) likelihood ratios with 95% confidence intervals (CIs). We extracted aspects of study design, characteristics of the population, index test, reference standard, and type of vertebral fracture. Meta-analysis was not possible due to heterogeneity of studies and index tests, therefore the analysis was descriptive. We calculated sensitivity, specificity, and LRs for each test and used these as an indication of clinical usefulness. Two review authors independently conducted risk of bias and applicability assessment using the QUADAS-2 tool.

Main results

This review is an update of a previous Cochrane Review of red flags to screen for vertebral fracture in people with low back pain. We included 14 studies in this review, six based in primary care, five in secondary care, and three in tertiary care. Four studies reported on 'osteoporotic vertebral fractures', two studies reported on 'vertebral compression fracture', one study reported on 'osteoporotic and traumatic vertebral fracture', two studies reported on 'vertebral stress fracture', and five studies reported on 'unspecified vertebral fracture'. Risk of bias was only rated as low in one study for the domains reference standard and flow and timing. The domain patient selection had three studies and the domain index test had six studies rated at low risk of bias. Meta-analysis was not possible due to heterogeneity of the data. Results from single studies suggest only a small number of the red flags investigated may be informative.

In the primary healthcare setting, results from single studies suggest 'trauma' demonstrated informative +LRs (range: 1.93 to 12.85) for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture' (+LR: 6.42, 95% CI 2.94 to 14.02). Results from single studies suggest 'older age' demonstrated informative +LRs for studies in primary care for 'unspecified vertebral fracture' (older age greater than 70 years: 11.19, 95% CI 5.33 to 23.51). Results from single studies suggest 'corticosteroid use' may be an informative red flag in primary care for 'unspecified vertebral fracture' (+LR range: 3.97, 95% CI 0.20 to 79.15 to 48.50, 95% CI 11.48 to 204.98) and 'osteoporotic vertebral fracture' (+LR: 2.46, 95% CI 1.13 to 5.34); however, diagnostic values varied and CIs were imprecise. Results from a single study suggest red flags as part of a combination of index tests such as 'older age and female gender' in primary care demonstrated informative +LRs for 'unspecified vertebral fracture' (16.17, 95% CI 4.47 to 58.43).

In the secondary healthcare setting, results from a single study suggest 'trauma' demonstrated informative +LRs for 'unspecified vertebral fracture' (+LR: 2.18, 95% CI 1.86 to 2.54) and 'older age' demonstrated informative +LRs for 'osteoporotic vertebral fracture' (older age greater than 75 years: 2.51, 95% CI 1.48 to 4.27). Results from a single study suggest red flags as part of a combination of index tests such as 'older age and trauma' in secondary care demonstrated informative +LRs for 'unspecified vertebral fracture' (+LR: 4.35, 95% CI 2.92 to 6.48). Results from a single study suggest when '4 of 5 tests' were positive in secondary care, they demonstrated informative +LRs for 'osteoporotic vertebral fracture' (+LR: 9.62, 95% CI 5.88 to 15.73).

In the tertiary care setting, results from a single study suggest 'presence of contusion/abrasion' was informative for 'vertebral compression fracture' (+LR: 31.09, 95% CI 18.25 to 52.96).

Authors' conclusions

The available evidence suggests that only a few red flags are potentially useful in guiding clinical decisions to further investigate people suspected to have a vertebral fracture. Most red flags were not useful as screening tools to identify vertebral fracture in people with low back pain. In primary care, 'older age' was informative for 'unspecified vertebral fracture', and 'trauma' and 'corticosteroid use' were both informative for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture'. In secondary care, 'older age' was informative for 'osteoporotic vertebral fracture' and 'trauma' was informative for 'unspecified vertebral fracture'. In tertiary care, 'presence of contusion/abrasion' was informative for 'vertebral compression fracture'. Combinations of red flags were also informative and may be more useful than individual tests alone. Unfortunately, the challenge to provide clear guidance on which red flags should be used routinely in clinical practice remains. Further research with primary studies is needed to improve and consolidate our current recommendations for screening for vertebral fractures to guide clinical care.

PLAIN LANGUAGE SUMMARY

Use of red flags to screen for vertebral fractures in people with low back pain

Key messages

– The four best red flags for vertebral fractures in people with low back pain were corticosteroid use (e.g. medicines that can weaken bones), older age (e.g. aged above 70 years), trauma (e.g. a fall), and a contusion (bruising) or abrasion (cuts and grazes).

– More research is needed to identify the best red flags or combination of red flags to screen for spinal fractures.

Red flags to screen for spinal fractures

Red flags for spinal fractures are signs and symptoms found by a health professional (e.g. doctor, physiotherapist) during an examination that warn that something is wrong within the spine (backbone). The accuracy of red flags is important, as low-quality tests can lead to

incorrect diagnosis and treatment. On the one hand, if the tests are not accurate, people without a spinal fracture (break) may undergo unnecessary imaging (e.g. X-ray, magnetic resonance imaging that uses radio waves to produce detailed images of the inside of the body). Some of these imaging methods lead to radiation exposure, extra costs, and added worry for the patient. On the other hand, missing a spinal fracture will result in a delay in receiving treatment and reduce quality of life. Therefore, identifying the most accurate red flags to screen for spinal fractures is needed.

What did we want to find out?

We wanted to assess how accurate the red flags used to screen for spinal fracture are in people presenting with low back pain. Where possible, we reported results of red flags separately for the different types of spinal fracture, such as osteoporotic vertebral compression fractures (e.g. fractures due to osteoporosis), vertebral traumatic fracture (e.g. due to falls), vertebral stress fracture (e.g. rapid increase in load on the spine), or unspecified vertebral fracture (e.g. no specific cause reported).

What did we do?

We updated a previous Cochrane Review. We searched for studies that investigated the accuracy of red flags across different healthcare settings. We included studies that compared results of history taking and physical examination (or both) (known as index tests or red flags) with different types of imaging (known as the reference standard) to identify spinal fractures in people with low back pain. We also included studies if they reported on the results of red flags separately for the different types of spinal fractures.

What did we find?

Fourteen studies investigated different red flags used to identify spinal fractures, and most of the red flags were not accurate or useful. Overall, the four best red flags found were corticosteroid use (e.g. medicines that can weaken bones), person's age (e.g. aged above 70 years), trauma (e.g. a fall), and a contusion (bruising) or abrasion (cuts and grazes).

In the primary healthcare setting (e.g. general practitioners), 'trauma' as a red flag was best to screen for 'unspecified spinal fracture' and 'osteoporotic spinal fracture'. 'Older age' as a red flag was best to screen for 'unspecified spinal fracture' in primary care. 'Corticosteroid use' may be useful as a red flag in primary care to screen for 'unspecified spinal fracture' and 'osteoporotic spinal fracture'. Red flags as part of a combination of index tests such as 'older age and female gender' as a red flag in primary care is best to screen for 'unspecified spinal fracture'.

In the secondary healthcare setting (e.g. specialists and consultants), 'trauma' as a red flag is best to screen for 'unspecified spinal fracture' and 'older age' for 'osteoporotic spinal fracture'. Red flags as part of a combination of index tests such as 'older age and trauma' in secondary care as a red flag is best to screen for 'unspecified spinal fracture'. When 'four of five tests' are positive in secondary care as a red flag, it may be used to screen for 'osteoporotic spinal fracture'.

In the tertiary care setting (e.g. specialised care in a hospital setting), the 'presence of contusion/abrasion' as a red flag was best to screen for 'spinal compression fracture'.

What are the limitations of the evidence?

A limitation of our review is that most of the included studies were different in terms of the healthcare setting they were performed in, used different study designs, or presented data for different types of spinal fracture, which made drawing conclusions difficult. Many of the red flags investigated were from single studies and few studies investigated the same index tests. There was also little uniform agreement on the definition of red flags (e.g. for corticosteroids, it was not clear how long and how much was used), which may explain why the accuracy of some red flags varied from study to study. Some red flags may also affect different types of spinal fractures differently; however, this was not clearly reported in most cases.

How up to date is this evidence?

The evidence is up to date to July 2022.

SUMMARY OF FINDINGS

Summary of findings 1. Results of potentially useful red flags

Review question: what is the accuracy of red flags to screen for vertebral fracture in people presenting with low back pain or for lumbar examination?

Patient population: people with low back pain or requiring examination of the lumbar spine when presenting to care in primary, secondary, or tertiary settings

Index tests: all relevant features taken during a history or physical examination

Target condition: vertebral fracture

Reference standard: diagnostic imaging (MRI, CT, X-ray, bone scan)

Included studies: 4 prospective cohorts, 4 retrospective chart reviews

Main limitations: small number of studies included; large heterogeneity between studies and index tests precluded pooling of results; descriptive analysis presented; inadequate reporting of methods

Population	Reference standard	Number of studies with high risk of bias per QUADAS-2 domain: participant selection/index test/ reference standard/flow and timing	Sensitivity (95% CI) *or range of estimates*[a]	Specificity (95% CI) *or range of estimates*[a]	+LR (95% CI)	−LR (95% CI)
Results: single red flags						
Primary care: 1 study **No. of participants:** 1172 **Index test:** age > 70 years	Follow-up with diagnostic imaging	0/0/0/1	0.50 (0.16 to 0.84)	0.96 (0.94 to 0.97)	11.19 (5.33 to 23.51)	0.52 (0.26 to 1.05)
Primary care: 2 studies **No. of participants:** 3179 **Index test:** age > 74 years	X-ray, follow-up with diagnostic imaging	0/0/0/1	Range: 0.25–0.59	Range: 0.84–0.97	Range: 3.69–9.39	Range: 0.49–0.77

Primary care: 3 studies **No. of participants:** 2152 **Index test:** history of corticosteroid use	X-ray, follow-up with diagnostic imaging	2/0/2/2	Range: 0.00–0.25	Range: 0.93–0.99	Range: 2.46–48.50	Range: 0.75–0.98
Primary care: 4 studies **No. of participants:** 3023 **Index test:** trauma	X-ray, follow-up with diagnostic imaging	2/0/2/2	Range: 0.21–0.65	Range: 0.89–0.98	Range: 1.93–12.85	Range: 0.37–0.88
Secondary care: 1 study **No. of participants:** 9940 **Index test:** age > 70 years	Imaging[b]	0/0/0/0	0.31 (0.27 to 0.35)	0.80 (0.79 to 0.81)	1.55 (1.36 to 1.76)	0.86 (0.82 to 0.91)
Secondary care: 1 study **No. of participants:** 9940 **Index test:** trauma	X-ray	0/0/0/0	0.25 (0.21 to 0.29)	0.89 (0.88 to 0.89)	2.18 (1.86 to 2.54)	0.85 (0.81 to 0.89)
Tertiary care: 3 studies **No. of participants:** 1142 **Index test:** trauma	X-ray	0/1/0/0	Range: 0.07–1.00	Range: 0.51–0.60	Range: 0.18–1.93	Range: 0.12–1.54
Tertiary care: 1 study **No. of participants:** 552 **Index test:** contusion/abrasion	X-ray	0/0/0/0	0.85 (0.70 to 0.94)	0.97 (0.95 to 0.98)	31.09 (18.25 to 52.96)	0.15 (0.07 to 0.32)
Results: combined red flags						
Primary care: 2 studies **No. of participants:** 3182 **Index test:** female and age > 64 years	X-ray, follow-up with diagnostic imaging	0/0/0/1	Range: 0.59–0.63	Range: 0.79–0.96	Range: 2.75–14.58	Range: 0.39–0.52
Primary care: 2 studies **No. of participants:** 3280	X-ray, follow-up with diagnostic imaging	0/0/0/1	Range: 0.25–0.45	Range: 0.90–0.98	Range: 4.36–16.17	Range: 0.62–0.76

Index test: female and age > 74 years						
Primary care: 1 study **No. of participants:** 1172 **Index test:** 2/4 red flags positive (Henschke)[c]	Follow-up with diagnostic imaging	0/0/0/1	0.63 (0.24 to 0.91)	0.96 (0.95 to 0.97)	15.48 (8.45 to 28.36)	0.39 (0.16 to 0.96)
Secondary care: 1 study **No. of participants:** 9940 **Index test:** age > 70 and Trauma	Imaging[b]	0/0/0/0	0.05 (not reported)	0.99 (not reported)	4.35 (2.92 to 6.48)	0.96 (0.94 to 0.98)
Secondary care: 1 study **No. of participants:** 1448 **Index test:** 4/5 features positive (Roman)[d]	X-ray or CT	0/0/0/1	0.37 (0.22 to 0.54)	0.96 (0.95 to 0.97)	9.62 (5.88 to 15.73)	0.66 (0.52 to 0.84)
Tertiary care: 1 study **No. of participants:** 108 **Index test:** trauma and neurological signs	X-ray	0/0/0/0	0.29 (0.04 to 0.71)	0.98 (0.93 to 1.00)	14.43 (2.38 to 87.64)	0.73 (0.46 to 1.15)

CI: confidence interval; **CT:** computed tomography; **LR:** likelihood ratio; **MRI:** magnetic resonance imaging.

[a] Sensitivity and specificity are the point estimate and 95% CI when only one study investigated an index test and the point estimate range when multiple studies investigated the same test.
[b] Imaging type not reported.
[c] Henschke combination: age > 70 years, significant trauma, prolonged corticosteroid use, altered sensation from the trunk down.
[d] Roman combination: age > 52 years, no leg pain present, body mass index < 23, does not exercise regularly, female gender.

BACKGROUND

Low back pain continues to be the leading cause for years lived with disability according to the Global Burden of Disease Study (Wu 2020). It is estimated that 577 million people experience low back pain globally, and in 2017 the estimated global years lived with disability was 64.9 million (Wu 2020). It is widely agreed that low back pain can be very disabling and impose an enormous social and economic burden on the community (Buchbinder 2018). Despite the disability burden of low back pain, it is rare for presentations to be due to a serious underlying condition (e.g. cauda equina syndrome, spinal tumours, and cancer) (Hartvigsen 2018). Most presentations can be managed in primary care without further diagnostic investigation (Oliveira 2018). However, there is an increasing number of low back pain presentations to emergency departments (Ferreira 2021).

Guidelines recommend that a thorough history and clinical assessment be used to identify people with suspicion of serious pathology who then require further diagnostic work-up to confirm the diagnosis. In the primary care setting, between 1% and 5% of all people who present with low back pain will have a serious spinal pathology (Downie 2014; Hartvigsen 2018; Henschke 2009; Williams 2013). The prevalence of serious pathology in people presenting to the emergency department is thought to be somewhat higher, one review reporting figures between 2.5% and 5.1% (Galliker 2019). In both primary care and emergency department settings vertebral fractures are amongst the most common serious pathologies which a clinician needs to screen for.

The use of red flags with high false-positive rates may contribute to unnecessary imaging and costs in low back pain management (Williams 2013). In contrast, red flags with high false-negative rates can lead to failure to recognise serious pathology, resulting in delayed testing and treatment, and may increase patient morbidity (DePalma 2020).

Target condition being diagnosed

Lumbar vertebral fractures can include vertebral traumatic fractures, vertebral stress fractures, and acute osteoporotic vertebral compression fractures. These three types of vertebral fractures may have different distinguishing clinical histories and clinical features, and may require different diagnostic work-up and management.

Lumbar vertebral stress fractures are thought to be secondary to chronic low-grade trauma or repetitive loading, which may occur in sport or training of new military recruits (Alqarni 2015). Lumbar vertebral traumatic fracture generally results from high-energy trauma (Lee 2016; Marongiu 2018). Osteoporotic vertebral compression fractures can occur from minor trauma (e.g. fall from a standing height or less) or without any definitive trauma (e.g. non-acute fracture) in older people with osteoporosis (Lee 2016). Osteoporotic vertebral fractures are associated with a two- to eight-fold risk of mortality (Schousboe 2016), and a five-fold increased risk of sustaining a further vertebral fracture (Marongiu 2018), and future hip fracture (Warriner 2011). One systematic review of 14 primary care studies (14,860 adults aged 50 years and older) found the median prevalence of vertebral fractures in people presenting with low back pain was estimated at 3.6% in primary care and 6.5% in secondary and tertiary care (Downie 2014).

Index test(s)

The purpose of the index test is screening for lumbar vertebral fracture. In people presenting with low back pain, clinical practice guidelines generally recommend assessing for the following red flags to help identify people with a greater suspicion of vertebral fracture: major/significant trauma, use of corticosteroids, greater than 50 years of age, and obvious structural deformity (Verhagen 2016).

Despite the inclusion of several red flags in guidelines, the usefulness of red flags to screen for serious pathology, such as fracture, continues to be debated (Grunau 2017; Underwood 2009). Several red flags endorsed by previous guidelines have low or untested diagnostic accuracy with little information on their utility for clinical practice (e.g. dose and duration of corticosteroid use not reported) (Cook 2018), or are based on evidence from single studies only (Parreira 2019).

In 2013, our group published a Cochrane Review of eight studies which evaluated 29 different index tests used to screen people presenting with low back pain for vertebral fracture (Williams 2013). Results from individual studies suggested many of the endorsed red flags, such as thoracic pain and neurological deficit, were uninformative because these red flags had both positive (+LR) and negative (−LR) likelihood ratios very close to 1 (Williams 2013). A +LR or −LR close to 1 indicates that the test does not substantially increase or decrease the likelihood of a fracture being present (if positive) or absent (if negative) (Deeks 2004). There was also variation between studies in the estimates of the diagnostic accuracy of the index tests. For example, for significant trauma, the sensitivity ranged from 0.25 (95% confidence interval (CI) 0.03 to 0.65) to 0.65 (95% CI 0.44 to 0.83) and the specificity ranged from 0.90 (95% CI 0.86 to 0.93) to 0.98 (95% CI 0.96 to 0.98). For corticosteroid use, the sensitivity ranged from 0.00 (95% CI 0.00 to 0.23) to 0.25 (95% CI 0.03 to 0.65), however the test was highly specific (0.99, 95% CI 0.98 to 1.00). For older age (greater than 74 years), the sensitivity ranged from 0.25 to 0.59 and the specificity ranged from 0.84 to 0.97. The +LRs for the different red flags were: significant trauma (from 3.42 (95% CI 1.57 to 7.45) to 12.85 (95% CI 8.58 to 19.24)); older age (from 3.69 (95% CI 3.00 to 4.53) to 9.39 (95% CI 2.69 to 32.75)); and corticosteroid use (from 3.97 (95% CI 0.20 to 79.15) to 48.50 (95% CI 11.46 to 204.98)).

Red flags used in isolation had only modest diagnostic accuracy. Combinations of red flags seemed more informative than red flags used alone. For example, age greater than 64 years had a +LR of 7.13 which increased to 16.17 when combined with female gender (Williams 2013). However, when combining red flags, the imprecision of the estimates also increased. One red flag in a tertiary care setting appeared informative with a high +LR for contusion/abrasion (31.09, 95% CI 18.25 to 52.96). Overall, there was uncertainty about the value of most red flags, as estimates were imprecise, there was a lack of agreement for older age cut-off points between studies and an overall lack of uniform agreement of definitions of index tests, and most red flags were only evaluated within a single study.

Clinical pathway

Most clinical practice guidelines recommend history and physical examination to identify people with a higher likelihood of low back pain due to vertebral fracture (Oliveira 2018), who require

further diagnostic work-up. Different types of vertebral fractures may require different imaging procedures.

The American College of Radiology (ACR) recommends standard radiography for initial imaging in people presenting with low back pain with a suspicion of osteoporotic vertebral compression fracture (Patel 2016). Genant's semi-quantitative assessment and the algorithm-based qualitative approach are methods used to classify and confirm the presence of a vertebral fracture on standard radiographs (Genant 1993; Jiang 2004). Computed tomography (CT) is recommended over a plain radiograph if there is suspicion of vertebral traumatic fracture and bone scan with single-photon emission computerised tomography (SPECT) if there is suspicion of vertebral stress fracture (Patel 2016).

Due to concerns of radiation exposure from standard radiography and CT (Berrington de González 2009; Wall 2006), vertebral fracture assessment by dual-energy X-ray absorptiometry (DXA) is a potential alternative to assess for osteoporotic and traumatic vertebral fractures (Lee 2016). However, there is currently insufficient evidence to replace standard radiography with DXA or CT due to its reduced ability to detect lower-grade vertebral fractures (Fuerst 2009; Lee 2016; Lems 2021), reduced image resolution (Bazzocchi 2012; Damiano 2006; Schousboe 2006), and numerous potential sources of error (Johansson 2020).

Magnetic resonance imaging (MRI) is now preferred over CT in evaluating people with suspicion of lumbar stress fracture. MRI can detect bone marrow oedema early, facilitating early diagnosis of lumbar stress reaction and fracture without radiation exposure (Ang 2016; Astur 2015). Given that, most people presenting with low back pain with suspicion of lumbar stress fracture tend to be from younger age groups, MRI may be beneficial as first-line imaging given its high sensitivity and specificity, and non-ionising radiation (Alqarni 2015; Astur 2015; Cheung 2018; Murthy 2012). However, the higher cost of MRI and potential increase in downstream expenditure need to be considered (Jarvik 2003).

Prior test(s)

In this review on index tests to screen for vertebral fracture, no prior tests are required as the index tests are used in the initial screening to increase suspicion of vertebral fracture in people presenting with low back pain.

Role of index test(s)

Index tests are used as screening tests, that is to rule in vertebral fracture in order to identify those who may require further diagnostic testing and specialised care. However, index tests are also used as screening tests to rule out vertebral fracture in order to reduce unnecessary imaging, treatments, and costs.

Alternative test(s)

Apart from a combination of history taking and physical examination, there are no alternative screening tests for vertebral fracture in people presenting with low back pain. Routinely imaging all people with low back pain would overburden healthcare systems and is discouraged in current guidelines (Oliveira 2018).

Rationale

In light of the publication of primary diagnostic studies (de Schepper 2016; Premkumar 2018), and due to the evolving guidance for the most appropriate methods to systematically review studies of diagnostic test accuracy (Deeks 2022; Whiting 2011), we aimed to update the previous Cochrane Review (Williams 2013). The results of this review will provide researchers and clinicians with more clarity in the use of red flags when screening for vertebral fracture in people presenting with low back pain in primary, secondary, and tertiary care settings.

OBJECTIVES

To assess the diagnostic accuracy of red flags used to screen for vertebral fracture in people presenting with low back pain. Where possible, we reported results of red flags separately for different types of vertebral fracture (i.e. acute osteoporotic vertebral compression fracture, vertebral traumatic fracture, vertebral stress fracture, unspecified vertebral fracture).

Secondary objectives

To assess the influence of sources of heterogeneity on the diagnostic accuracy of red flag tests, if appropriate. In particular, if possible, we explored heterogeneity within the different types of vertebral fracture (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, and vertebral stress fracture) due to factors such as healthcare setting (e.g. primary, emergency, or secondary care). We also aimed to assess the mean age of participants in each included study at different cut-off thresholds (e.g. greater than 50 years of age versus greater than 70 years of age) across different studies, as this has previously been shown to influence results.

METHODS

Criteria for considering studies for this review

Types of studies

We considered primary diagnostic studies if they compared the results of history taking and physical examination (or both) findings (index test) with a reference standard test for the identification of vertebral fracture in people presenting with low back pain. We included studies that reported results of red flags separately for the different types of vertebral fracture or studies that reported combined results for vertebral fractures. We included studies of cross-sectional designs that presented sufficient data to allow estimates of diagnostic accuracy (such as sensitivity and specificity) to be derived. We contacted authors of abstracts and conference proceedings for the full text; if the full text was unavailable, we excluded such studies.

We excluded studies with participants diagnosed with pathological vertebral fractures (e.g. malignant fractures). We excluded case-control studies (i.e. two-gate studies) that had included some participants (cases) already diagnosed with vertebral fracture and some participants (controls) without vertebral fracture as these types of studies may overestimate both sensitivity and specificity (Whiting 2013).

We retrieved and included studies published in any language. If necessary, we organised appropriate translation of potentially eligible articles.

Participants

We included studies if they evaluated adolescents (from the age of 15 years and over) or adults presenting to primary, secondary, or tertiary care for management of low back pain. We excluded paediatric populations as presentation, risk factors, and fracture types would likely differ compared to adult populations.

Index tests

Studies evaluating any aspects of the history or physical examination (or both) for people presenting with low back pain were eligible for inclusion. This included all information regarding:

- demographic characteristics (e.g. age, gender);
- clinical and medical history (e.g. pain intensity, previous history of falls, previous or current medication use);
- physical examination results (e.g. tenderness on palpation, muscle strength testing, or lumbar range of motion).

We included studies if the diagnostic accuracy of individual red flags was evaluated in isolation, or as part of a combination (e.g. female sex, greater than 70 years of age, significant trauma, and prolonged corticosteroid use). We extracted or derived indices of diagnostic performance from data presented in each primary study for each red flag or combination of red flags. We excluded studies that used only a 'clinical diagnosis' or 'global clinician judgement' (without specifying which diagnostic features were considered) as the diagnostic accuracy of clinical judgement based on undefined parameters is difficult to interpret.

Target conditions

We included studies that investigated people presenting with low back pain or for lumbar examination and reported appropriate data to calculate an estimate of the diagnostic accuracy of tests for vertebral fracture. Where possible, we described results separately for different types of vertebral fracture (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, and vertebral stress fracture).

Where possible, we attempted to categorise the different types of vertebral fractures under the following definitions.

- Acute osteoporotic vertebral compression fracture: insufficiency fractures (i.e. a fracture due to normal stresses on an abnormal bone).
- Vertebral traumatic fracture: any vertebral fracture due to a trauma mechanism in normal bone (e.g. direct force to the lumbar spine).
- Vertebral stress fracture: fractures as a result of repetitive trauma over time (i.e. a fracture due to abnormal stresses on a normal bone).
- Unspecified vertebral fracture: if the type of fractures was not specified we referred to this as unspecified vertebral fracture.

If it was not possible to differentiate between the different types of vertebral fractures in the included studies, then we categorised these fractures as 'unspecified vertebral fractures'.

Reference standards

If studies reported sufficient data, we explored how the different reference standards impacted the diagnostic accuracy of the index test. For traumatic vertebral fractures, CT was used as the preferred reference test. If CT scan was not available, we used MRI as the reference test. If neither was used, we used whichever imaging modality was presented as the reference standard (e.g. X-ray). For osteoporotic vertebral fractures and vertebral stress fractures, we used MRI as the preferred reference test. If MRI was not available, we used bone scintigraphy with SPECT as the reference test. If neither was used, we used whichever imaging modality was presented as the reference standard (e.g. X-ray). However, we anticipated that most studies would not separate reference standards based on the different types of vertebral fracture.

We also considered clinical follow-up after the initial consultation as a weaker reference standard if suspected vertebral fracture was subsequently confirmed by imaging.

Search methods for identification of studies

Electronic searches

The search strategy was based upon the Williams 2013 Cochrane Review and was updated in collaboration with a librarian. Williams 2013 search was run from inception to 7 March 2012. We searched relevant databases for eligible diagnostic studies from January 2012 to 26 July 2022. We searched the following databases.

- MEDLINE (Ovid, from 1946) (Appendix 1)
- Embase (Embase.com; Appendix 2)
- CINAHL (EBSCO; Appendix 3)

Searching other resources

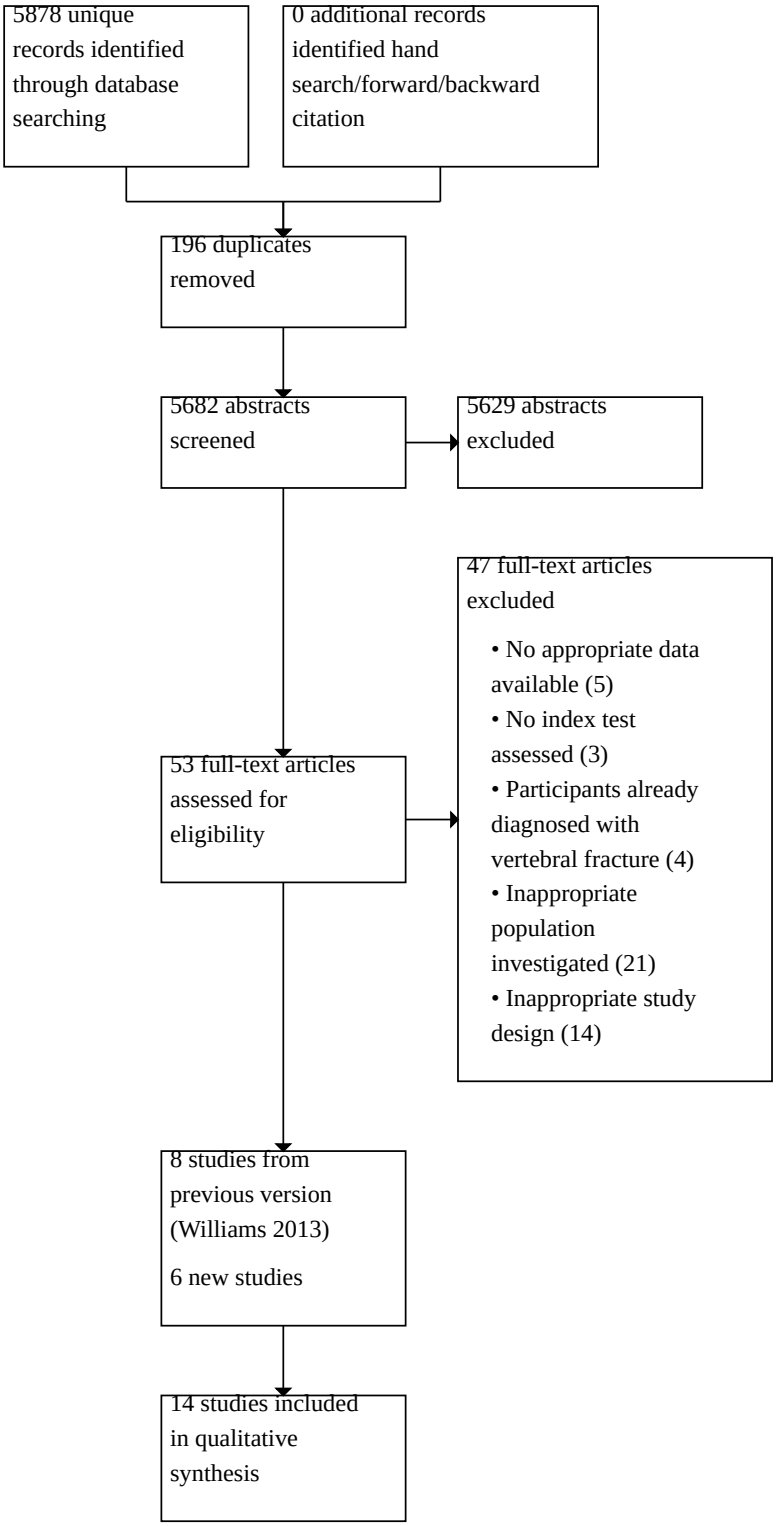
We checked the reference lists of all included publications. We performed forward citation tracking of the included studies through Web of Science, Scopus, and Google Scholar. We sent the final list of included studies for review by experts in the field to minimise the risk of missing any relevant studies.

Data collection and analysis

Selection of studies

Two review authors (CSH and QC) independently applied the selection criteria to all studies (titles and abstracts) identified from the literature search. The two review authors (CSH and QC) independently excluded clearly irrelevant studies. Two review authors (CSH and QC) then retrieved and independently assessed the full-text publications using the predefined inclusion criteria. We resolved all disagreements through consensus, and consulted other review authors where necessary. We contacted relevant study authors via email if any study information needed to be clarified. All excluded studies are presented with reasons for their exclusion in the Characteristics of excluded studies table and the study selection process is presented in the PRISMA-DTA diagram (McInnes 2018) (Figure 1).

Figure 1. Study flow diagram.



Data extraction and management

Two review authors (CSH and QC) independently extracted the relevant data from the included studies. This included data on the characteristics of the included studies such as:

- participant characteristics;
- type of vertebral fracture;
- types of index tests and reference standards, and their descriptors;
- relevant aspects of the study methods;
- data for the assessment of diagnostic accuracy.

Characteristics of participants (and studies) included details on:

- healthcare setting (e.g. primary care (medical centre, physiotherapist, chiropractor), emergency departments; secondary care (outpatient hospital service, medical hospital services, specialist services); tertiary care (higher specialised care in hospital setting));
- inclusion and exclusion criteria;
- enrolment procedures (consecutive or non-consecutive);
- number of participants (including number eligible and enroled in the study);
- number of participants receiving the index test and reference standard;
- number of participants for whom results were reported in the two-by-two table;
- reasons for withdrawal;
- participant demographics (age, gender, duration, and history of low back pain).

Test characteristics included details on:

- type of index test;
- methods of execution of index tests;
- threshold for positive result on an index test;
- experience and expertise of the assessors;
- type of reference standard;
- cut-off points for diagnosing vertebral fracture on a reference standard (e.g. quantitative radiographic measures) where relevant.

Thresholds for 'positive' results on an index test may vary across studies and some studies may present the diagnostic performance of an index test at several different cut-off points (e.g. age greater than 50 years versus age greater than 70 years). This is presented on a Receiver Operating Characteristic (ROC) curve. We extracted diagnostic accuracy data for all cut-off points from the included studies.

To assess diagnostic accuracy, we extracted data for diagnostic two-by-two tables (true positive, false positive, true negative, and false negative counts) from the publications or reconstructed them using information from other relevant parameters (sensitivity, specificity, or predictive values). If the reported sensitivity, specificity, and positive and negative predictive values did not match the reported two-by-two table data we reported results based on the raw two-by-two data. However, if there were large discrepancies in the values calculated from raw two-by-two data and the values presented in a study, we contacted the study authors to seek clarification of the data. If extracting raw data to produce a two-by-two table was not possible, we used the sensitivity and specificity values reported. We discussed any discrepancies in calculated two-by two data and values presented in a study with the author team and reported them in the review. For eligible studies where we could not reconstruct the diagnostic two-by-two tables, we presented them in the review but did not include them in the quantitative analysis. Two review authors (CSH and QC) independently extracted the data to ensure adequate reliability of collected data. For each study, where possible, we presented in tables aspects of study design, characteristics of the population, index test, and reference standards, and type of vertebral fracture. If a review author was also an author of one of the primary diagnostic studies, they were not involved in the data extraction or quality rating of that study.

Assessment of methodological quality

Two review authors (CSH and QC) independently assessed the methodological quality of each included study using the QUADAS-2 checklist (Whiting 2011), as recommended by the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Appendix 4; Deeks 2022). The QUADAS-2 checklist assessed study quality for all studies included in the previous review and the additional studies included in this updated review. We tailored the QUADAS-2 tool to our review and made one change by removing "was a case-control design avoided?" as case control studies were excluded. The QUADAS-2 checklist consists of four domains: patient selection, index test, reference standard, and flow and timing. We assessed all four domains for risk of bias and all three domains for concerns regarding applicability.

The same review authors classified each item as 'yes' (adequately addressed); 'no' (inadequately addressed); or 'unclear' (inadequate detail presented to allow a judgement to be made). We resolved disagreements through discussion and if necessary, consulted a third review author (CGM or MJH).

Statistical analysis and data synthesis

Statistical analysis followed the approaches outlined in Chapter 10 of the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Deeks 2022). We extracted indices of diagnostic performance or derived them from data presented in each primary study for each red flag or combination of red flags. Sensitivity, specificity, and likelihood ratios (LR) of the index tests were the primary measures of diagnostic test accuracy, both in the text and Summary of findings 1. Summary of findings 1 includes the following variables, but was not limited to: the review question, patient population, index test, target condition, reference standard, study quality, included studies characteristics, sensitivity, specificity, and LRs. We generated diagnostic two-by-two tables, from which we calculated and presented sensitivities and specificities for each index test with 95% CIs and present these in forest plots. We calculated LRs for each test and used these as an indication of clinical informativeness. Likelihood ratios represent the increase (with a positive test) or decrease (with a negative test) in likelihood of a person having a vertebral fracture (Deeks 2004). A +LR greater than 1 indicates that a positive test is associated with an increase in the likelihood of a vertebral fracture being present. A negative likelihood ratio (−LR) less than 1 indicates that a negative test is associated with a decrease in the likelihood of a vertebral fracture (Deeks 2004). Criteria for clinical informativeness

of LRs are presented in Table 1. In addition, we presented an ROC plot of sensitivity versus 1 − specificity to display the data. If possible, we intended to present results of how different pretest probability cut-offs and positive or negative results impacted post-test probabilities.

We intended to use the bivariate random-effects model to compute summary estimates of sensitivity and specificity. The planned method accounts for both within-study and between-study variability, and allows for correlation that may exist between sensitivity and specificity across studies (Harbord 2007; Reitsma 2005). This approach would have assumed a binomial distribution within studies (Chu 2006), and we assumed the random effects to be normally distributed across studies. We intended to restrict these analyses to studies that showed sufficient clinical homogeneity (e.g. same index test and index test threshold, and a similar definition of vertebral fracture). Where possible, we intended to perform meta-analysis on estimates of sensitivity and specificity to calculate the +LR and −LR for each index test.

For index tests where the cut-off for test positivity varied across studies, we intended to use the Hierarchical Summary ROC (HSROC) method of Rutter and Gatsonis (Rutter 2001) to estimate a summary curve which we presented graphically if appropriate. Where a study had data for more than one cut-off point, we intended to use the most commonly reported cut-off point across studies. We initially intended to use SAS software for meta-analysis, however meta-analysis was not possible and instead used Review Manager 5 for all statistical analysis and data synthesis (Review Manager 2014).

Investigations of heterogeneity

Several factors can contribute to heterogeneity in the diagnostic accuracy of a test across multiple studies. Where possible, we investigated the potential influences of one study level variable at a time, such as healthcare setting (i.e. primary care, emergency department, secondary care), different types of vertebral fracture (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, and vertebral stress fracture), or different types of imaging. We chose healthcare setting as the prevalence of vertebral fracture may vary depending on the healthcare setting. We chose different types of vertebral fracture as the diagnostic accuracy of red flags may vary dependent on the type of fracture. We intended to perform meta-analysis on results separately for different healthcare settings and for each of the different types of vertebral fracture if data were available. If results for an index test were not reported separately for different types of fractures we reported the results as the diagnostic accuracy for 'unspecified vertebral fracture'.

Where feasible, we intended to include study-level covariates (e.g. different types of imaging) in the HSROC analysis models to investigate sources of heterogeneity. Where possible, we clearly acknowledged any unexplained heterogeneity between studies in Summary of findings 1, as per recommendations of the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Deeks 2022).

Sensitivity analyses

Where possible, we intended to perform a sensitivity analysis to investigate how individual QUADAS-2 key domains (i.e. patient selection, index test, reference standard, and flow and timing) affected accuracy estimates. That is, we aimed to assess whether estimates of sensitivity and specificity were affected if studies at high risk of bias in any of the QUADAS-2 domains were removed for the sensitivity analysis.

Assessment of reporting bias

Tests for publication bias in reviews of diagnostic test accuracy are not routinely used because study size may be related to test accuracy for reasons other than publication bias (Deeks 2005). Therefore, we did not report on this.

RESULTS

Results of the search

Our search identified 5878 records, After removing 196 duplicates, 5682 records remained. Of these, 5629 were deemed clearly irrelevant based on screening by title and abstract. There were only two relevant systematic reviews found during the updated search (Downie 2014; Galliker 2019). We reviewed the reference lists of both systematic reviews and performed forward citation searches on the reference lists of all included studies with no additional studies meeting the inclusion criteria. Of the 53 articles included for full-text review, we excluded 47 with reasons (five had no appropriate data available, three did not assess an index test, four had participants already being diagnosed with vertebral fracture, 21 investigated an inappropriate population, and 14 had an inappropriate study design) (see Characteristics of excluded studies table). Therefore, we included six new studies to add to the eight studies from the previous review (Williams 2013), resulting in 14 studies included for analysis.

Management of missing information

We contacted the authors of two studies for additional data or discrepancies (or both) (Premkumar 2018; Therriault 2020). Therriault 2020 provided some additional data but it was not suitable for our analyses. After contacting the authors again, they confirmed they no longer had access to the raw data we needed for our analysis to calculate the diagnostic accuracy for a combination of red flags (e.g. male sex, pain with lumbar extension, active pain worse than resting pain, or a combination of these). Therefore, we excluded this red flag from our analyses (Therriault 2020).

Description of studies

See Characteristics of included studies table.

Most studies were heterogeneous and did not present data based on the different vertebral fracture types. Four studies reported on 'osteoporotic vertebral fractures' (prevalence of fracture range; 3% to 68%) (Enthoven 2016; Jin 2020; Kilic 2021; Roman 2010), and two studies reported on 'vertebral compression fracture' (prevalence of fracture: Patrick 1983: 7.2%; Reinus 1998: 11%). One study reported on 'osteoporotic and traumatic vertebral fractures' (prevalence of fracture: 4.1%) but did not present data for the individual fracture types (van den Bosch 2004). Five studies did not differentiate between the different types of vertebral fracture and were considered to be reporting on 'unspecified vertebral fracture' (prevalence of fracture range: 0.68% to 6.5%) (Deyo 1986; Gibson 1992; Henschke 2009; Premkumar 2018; Scavone 1981). One study reported on 'early-stage spondylolysis', and defined spondylolysis as 'the pathogenesis of lumbar spondylolysis is considered to be a stress fracture' (prevalence of fracture: 52.5%)

(Sugiura 2021). Another study reported on 'active spondylolysis' and defined spondylolysis as a 'bone stress injury (stress reaction or fracture) of the pars interarticularis' (prevalence of fracture: 22%) (Therriault 2020). Based on each study's definition of spondylolysis, they were both categorised as 'vertebral stress fracture' (Sugiura 2021; Therriault 2020). The overall prevalence of vertebral fracture ranged from 0.68% to 68%; however, the patient sample in some studies was highly selective which likely explains the high prevalence of vertebral fracture. The high prevalence rate is explored further below in 'Findings based on healthcare setting'.

Six studies with 6365 participants were conducted in a primary healthcare setting (Deyo 1986; Enthoven 2016; Henschke 2009; Scavone 1981; Therriault 2020; van den Bosch 2004). The median sample size of the included studies in primary care was 948 (range: 621 to 2007). Of the six studies conducted in primary care, there were 86 potential red flags. One study presented data on a combination of red flags (Therriault 2020). Four studies used X-ray as the reference standard (Deyo 1986; Enthoven 2016; Scavone 1981; van den Bosch 2004), one study used different types of imaging methods (e.g. X-ray, MRI, CT, or a combination) based on the clinician's preference as the reference standard (Therriault 2020), and therefore not all participants received the same reference standard, and one study used clinical follow-up as a reference standard (Henschke 2009).

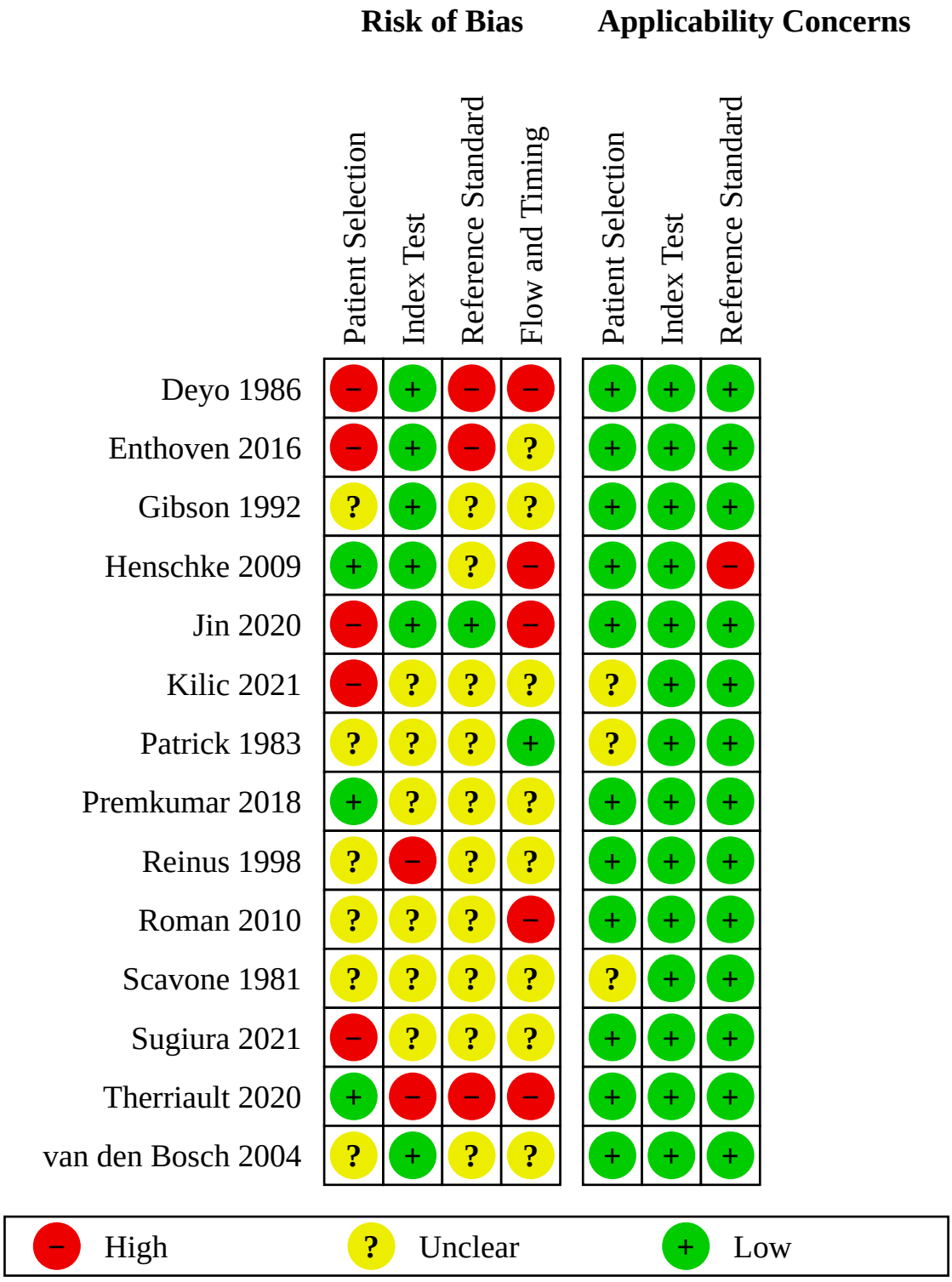
Five studies with 12,135 participants were conducted in a secondary healthcare setting (Jin 2020; Kilic 2021; Premkumar 2018; Roman 2010; Sugiura 2021). The median sample size of the included studies in secondary care was 510 (range: 101 to 9940). The five studies in secondary care identified 29 red flags. Two studies presented data on a combination of red flags (Premkumar 2018; Roman 2010). Two studies used X-ray as the reference standard (Kilic 2021; Premkumar 2018), two studies used MRI as the reference standard (Jin 2020; Sugiura 2021), and one study used X-ray or SPECT as the reference standard (Roman 2010).

Three studies with 1259 participants were conducted in a tertiary healthcare setting (Gibson 1992; Patrick 1983; Reinus 1998). The median sample size of the studies in tertiary care was 482 (range: 225 to 552). The three studies in tertiary care identified 17 red flags. One study presented data on a combination of red flags (Gibson 1992). All three studies used X-ray as the reference standard.

Methodological quality of included studies

Results of the quality assessment using the QUADAS-2 scale are presented in Figure 2. Further details on how judgements were made are presented in the Characteristics of included studies table. No studies demonstrated low risk of bias across all domains; however, 10/14 studies demonstrated low concerns regarding applicability.

Figure 2. Risk of bias and applicability concerns summary: review authors' judgements about each domain for each included study



Patient selection

For the domain of patient selection, only 3/14 studies demonstrated low risk of bias. Five of 14 studies demonstrated high risk of bias, due to participants being excluded for characteristics of their presentation that could be considered clinically relevant. For example, Jin 2020 and Sugiura 2021 excluded people with low back pain and pain radiating to the limbs which could be considered

part of the clinical presentation of people with vertebral fracture (Kim 2015). However, both studies investigated different types of vertebral fracture, Jin 2020 included people with 'osteoporotic vertebral fractures' and Sugiura 2021 included people with 'early-stage spondylolysis'. Enthoven 2016, Jin 2020, Kilic 2021, and Sugiura 2021 included specific age cut-offs as part of their inclusion criteria, potentially excluding people presenting with the target condition. The main concern in this domain was unclear risk of bias due to insufficient information regarding inappropriate exclusions in 6/14 studies. Eleven of 14 studies demonstrated low concerns regarding applicability and 3/14 studies demonstrated unclear concerns regarding applicability due to insufficient information available regarding patient presentation.

Index tests

For the domain of index test, 6/14 studies demonstrated low risk of bias. Two of 14 studies demonstrated high risk of bias, due to no prespecified thresholds being used for the index test (Reinus 1998), and index test results being interpreted with the knowledge of the reference standard result (Therriault 2020). Six of 14 studies demonstrated unclear risk of bias, as it was unclear if the index test results were interpreted with or without the knowledge of the reference standard result. All the included studies demonstrated low concerns for applicability.

Reference standards

For the domain of reference standard, only 1/14 studies demonstrated low risk of bias (Jin 2020). Three of 14 studies demonstrated high risk of bias as the reference standard result was interpreted with the knowledge of the index test result. Ten of 14 studies demonstrated unclear risk of bias, as it was unclear if the reference standard results were interpreted with or without the knowledge of the index test result. Thirteen of 14 included studies demonstrated low concerns regarding applicability. One study demonstrated high concern regarding applicability as they used clinical follow-up, and we considered this a weaker reference standard.

Flow and timing

For the domain of flow and timing, only 1/14 studies demonstrated low risk of bias (Patrick 1983). Five of 14 studies demonstrated high risk of bias, due to inappropriate interval time between index test and the reference standard (i.e. six weeks and greater) (Henschke 2009), not all participants included in the final analysis (i.e. 518/582 participants included for analysis) (Jin 2020), not all participants receiving the same reference test (Roman 2010; Therriault 2020), and one study for all three reasons listed (Deyo 1986). Eight of 14 included studies demonstrated unclear risk of bias, due to most studies not reporting clear interval times between index test and the reference standard.

Overall ratings

In summary, risk of bias was often rated unclear regarding patient selection, implementation of the index test and reference standard, and flow and timing. However, the applicability of the included studies did not demonstrate substantial concerns.

Findings

In total there were 130 potential index tests assessed across 14 studies. Studies used index tests assessing a similar construct but with slightly different wording/cut-offs (e.g. trauma = significant trauma, history of trauma, major trauma, direct trauma, or fall; older age = older age with eight different age cut-offs). There were 18 index tests which were presented as different combinations of red flags. Few index test descriptions were reported adequately for clinical use. We attempted to present index test results for the different types of vertebral fractures (i.e. osteoporotic vertebral compression fracture, vertebral traumatic fracture, vertebral stress fracture); however, this was not always possible as only some studies presented data based on the different types of fractures. Performing meta-analysis on estimates of test accuracy for this updated review was not possible due to the heterogeneity of the index tests presented across the studies, study healthcare setting, and different vertebral fracture types. Therefore, the findings of this review are presented descriptively.

Results for all tests are presented in Table 2, Table 3, Table 4, Table 5, and Table 6. Summary of findings 1 describes the performance of red flags that were potentially useful.

Findings based on healthcare setting

Primary care

Six studies were based in a primary healthcare setting and investigated 86 different index tests as potential red flags (Deyo 1986; Enthoven 2016; Henschke 2009; Scavone 1981; Therriault 2020; van den Bosch 2004). The potential red flags investigated included components of the history (e.g. age, gender, mechanism of injury) and physical examination (e.g. palpation, lumbar range of motion) and are presented below descriptively. Two studies reported on screening for 'osteoporotic vertebral fracture' (Enthoven 2016; van den Bosch 2004), one study reported on 'vertebral stress fracture' (Therriault 2020), and three studies reported results for 'unspecified vertebral fracture' (Deyo 1986; Henschke 2009; Scavone 1981). Four of the six studies included participants across a wide range of ages (0 to 98 years), however, Therriault 2020 included participants between the ages of 10 and 19 years and Enthoven 2016 only included adults over the age of 55 years. Any index tests investigated in Enthoven 2016 was in addition to the age cut-off used for inclusion to their study (i.e. over 55 years). For example, 'trauma' was investigated as a potential red flag, therefore the index test would be age over 55 years and 'trauma', not trauma alone.

Four studies investigated 'trauma' as a potential red flag (unspecified vertebral fracture: three studies, osteoporotic fracture: one study) (Deyo 1986; Enthoven 2016; Henschke 2009; Scavone 1981). The sensitivity for the index test 'trauma' for 'unspecified vertebral fracture' was generally low and varied greatly, ranging from 0.21 (95% CI 0.05 to 0.51) to 0.65 (95% CI 0.44 to 0.83) (Deyo 1986; Henschke 2009; Scavone 1981). However, 'trauma' was highly specific in all studies, ranging from 0.90 (95% CI 0.86 to 0.93) to 0.98 (95% CI 0.96 to 0.98). The +LRs ranged from 1.93 (95% CI 0.67 to 5.53) to 12.85 (95% CI 8.58 to 19.24) and the −LRs ranged from 0.36 (95% CI 0.22 to 0.62) to 0.88 (95% CI 0.67 to 1.17). Enthoven 2016 investigated 'trauma' as a red flag for 'osteoporotic vertebral fracture', but only included people over the age of 55 years; and demonstrated low sensitivity 0.21 (95% CI 0.09 to 0.39) but high specificity 0.97 (95% CI 0.95 to 0.98). The +LR 6.42 (95% CI 2.94 to 14.02) demonstrated informative diagnostic value, but uninformative −LR (0.82, 95% CI 0.68 to 0.97).

Four studies investigated 'older age' as a potential red flag (unspecified vertebral fracture: two studies, osteoporotic fracture: two studies), with eight different age cut-offs (Deyo 1986; Enthoven 2016; Henschke 2009; van den Bosch 2004). Diagnostic accuracy values for the different age cut-offs are presented in Table 2. The cut-off age of 'age greater than 70 years' demonstrated the highest diagnostic accuracy for 'unspecified vertebral fracture' with a sensitivity of 0.50 (95% CI 0.16 to 0.84), specificity of 0.96 (95% CI 0.94 to 0.97), +LR of 11.19 (95% CI 5.33 to 23.51), and −LR of 0.52 (95% CI 0.26 to 1.05) (Henschke 2009). One study investigated the 'older age' and 'gender' as part of a combination of red flags, which demonstrated higher diagnostic accuracy compared to older age alone. For example, when 'age greater than 74 years' and 'female gender' were combined, the +LR of a suspected 'unspecified vertebral fracture' increased from 9.39 (95% CI 2.69 to 32.75) to 16.17 (95% CI 4.47 to 58.43) (Henschke 2009).

Three studies investigated 'corticosteroid use' as a potential red flag (unspecified vertebral fracture: two studies, osteoporotic fracture: one study) (Deyo 1986; Enthoven 2016; Henschke 2009). The sensitivity for 'corticosteroid use' for 'unspecified vertebral fracture' was low and ranged from 0.00 (95% CI 0.00 to 0.23) to 0.25 (95% CI 0.03, 0.65), however demonstrated high specificity, ranging from 0.99 (95% CI 0.98 to 1.00) to 0.99 (95% CI 0.99 to 1.00) (Deyo 1986; Henschke 2009). The +LR varied greatly, ranging from 3.97 (95% CI 0.20 to 79.15) to 48.50 (95% CI 11.46 to 204.98), and the −LR ranged from 0.75 (95% CI 0.51 to 1.13) to 0.98 (95% CI 0.89 to 1.07). One study investigated 'corticosteroid use' as a red flag for 'osteoporotic vertebral fracture', but only included people over the age of 55 years (Enthoven 2016). It demonstrated low sensitivity (0.18, 95% CI 0.07 to 0.35) but high specificity (0.93, 95% CI 0.90 to 0.95). The +LR demonstrated informative diagnostic value (2.46, 95% CI 1.13 to 5.34), but uninformative −LR (0.88, 95% CI 0.75 to 1.04).

Two studies investigated a combination of index tests as red flags (unspecified vertebral fracture: one study, stress fracture: one study) (Henschke 2009; Therriault 2020). In Henschke 2009, the presence of at least two of four red flags (older age, gender, corticosteroid use, and history of trauma) for the suspicion of a 'unspecified vertebral fracture' demonstrated moderate sensitivity (0.63, 95% CI 0.24 to 0.91) and high specificity (0.96, 95% CI 0.95 to 0.97). Although there were large +LRs (15.48, 95% CI 8.45 to 28.36) when two of the four red flags were present, and for three of the four red flags present (906.11, 95% CI 50.37 to 16,299.10), the CIs were very wide and, therefore, these results should be interpreted with caution.

Secondary care

Five studies were based in a secondary healthcare setting and investigated 29 different index tests as potential red flags (Jin 2020; Kilic 2021; Premkumar 2018; Roman 2010; Sugiura 2021). These included components of the history (e.g. age, gender, mechanism of injury) and physical examination (e.g. lumbar range of motion) and are presented below descriptively. Three studies reported on screening for 'osteoporotic vertebral fracture' (Jin 2020; Kilic 2021; Roman 2010), one study reported on 'vertebral stress fracture' (Sugiura 2021), and one study reported on 'unspecified vertebral fracture' (Premkumar 2018). Three of five studies included age as part of their inclusion criterion and as an index test. Kilic 2021 only included geriatric adults over the age of 65 years. Sugiura 2021 only included adolescents under the age of 18 years. Index tests investigated in Sugiura 2021 would only apply to the age cut-off used for inclusion to their study. Jin 2020 only included postmenopausal women or women over 50 years old and men over 60 years old or people with a history of prolonged use of a glucocorticosteroid for three months or more. Participants with previous vertebral fractures were also included in this study (Jin 2020). Index tests investigated in Jin 2020 would apply to the participants who fit the 'age' cut-offs and 'corticosteroid use' used for inclusion to their study.

Three studies investigated 'older age' as a potential red flag (osteoporotic fracture: two studies, unspecified vertebral fracture: one study), with five different age cut-offs (Kilic 2021; Premkumar 2018; Roman 2010). Diagnostic accuracy values for the different age cut-offs are presented in Table 4. The cut-off age of 'age greater than 74 years' demonstrated the most useful diagnostic accuracy values for 'osteoporotic vertebral fracture' with a sensitivity of 0.53 (95% CI 0.41 to 0.64), specificity of 0.79 (95% CI 0.67 to 0.88), +LR of 2.51 (95% CI 1.48 to 4.27), and −LR of 0.60 (95% CI 0.46 to 0.79) (Kilic 2021). One study investigated 'trauma' as a potential red flag and demonstrated informative diagnostic accuracy values for 'unspecified vertebral fracture' with a sensitivity of 0.25 (95% CI 0.21 to 0.29), specificity of 0.89 (95% CI 0.88 to 0.89), +LR of 2.18 (95% CI 1.86 to 2.54), and −LR of 0.85 (95% CI 0.81 to 0.89) (Premkumar 2018). However, when 'older age' and 'trauma' were investigated as part of a combination of index tests for 'unspecified vertebral fracture', the diagnostic values increased compared to 'older age' or 'trauma' alone. For example, when 'older age greater than 70 years' and trauma were combined, the +LR of a suspected fracture increased from 1.55 (95% CI 1.36 to 1.76) to 4.35 (95% CI 2.92 to 6.48) (Premkumar 2018).

Two studies investigated individual components of the history (e.g. pain location, pain quality, etc.; Sugiura 2021) and physical examination (e.g. lumbar range of motion; Jin 2020; Sugiura 2021) as red flags (osteoporotic fracture; one study, stress fracture; one study). Jin 2020 investigated a series of dynamic movements (e.g. moving from lying to sitting) as a red flag for 'osteoporotic vertebral fracture'. This test demonstrated high sensitivity (0.99, 95% CI 0.97 to 1.00) and moderate specificity (0.68, 95% CI 0.60 to 0.75). The results of this index test should be interpreted with caution as the index test investigated dynamic movements is a combination with other 'index tests' used as part of their study inclusion criteria as mentioned above. The sample of participants in Jin 2020 was also highly selective which likely explains the high prevalence of vertebral fracture in this study (i.e. prevalence of vertebral fracture was 68%). In Sugiura 2021, the only index test with some diagnostic value was 'pain location' (i.e. pain on the left or right (or both) side or central pain) as a red flag for 'vertebral stress fracture'. This test demonstrated high sensitivity (0.89, 95% CI 0.77 to 0.96) and moderate specificity (0.63, 95% CI 0.48 to 0.77). The results of this index test should be interpreted with caution as the index test investigated (i.e. pain location) is in combination with 'age of 18 years or younger' which was used as part of their study inclusion criteria as mentioned above. The prevalence of a 'vertebral stress fracture' in Sugiura 2021 was also high (i.e. prevalence of spondylolysis was 53%), however previous research has shown this population sample is likely to have a high prevalence of spondylolysis due to their young age and participation in sporting activities (Selhorst 2019).

One study investigated eight different index tests as potential red flags for 'osteoporotic vertebral fracture' (Roman 2010); however, the diagnostic values of these tests varied greatly. Index tests with some diagnostic value included body mass index (BMI) less than 23, decreased pain on sitting, and no leg or buttock pain. The sensitivity of these three tests ranged from 0.29 (95% CI 0.15 to 0.46) to 0.37 (95% CI 0.22 to 0.54) and the specificity ranged from 0.81 (95% CI 0.79 to 0.83) to 0.86 (95% CI 0.84 to 0.88). These three index tests also demonstrated the largest +LRs; BMI less than 23 (2.22, 95% CI 1.44 to 3.42), decreased pain on sitting (1.56, 95% CI 0.94 to 2.59), and no leg or buttock pain (2.24, 95% CI 1.38 to 3.64). This study also investigated a combination of index tests (age greater than 52 years, no presence of leg pain, BMI less than 23, does not exercise regularly, and female gender) and found that having four of the five positive tests demonstrated a moderate increase in the likelihood of 'osteoporotic vertebral fracture' (+LR 9.62, 95% CI 5.88 to 15.73).

Tertiary care

Three studies were based in a tertiary healthcare setting and investigated 17 different index tests as potential red flags (Gibson 1992; Patrick 1983; Reinus 1998). The red flags investigated included components of the subjective history (e.g. mechanism of injury) and physical examination (e.g. palpation, neurological assessment) and are presented below descriptively. Two studies reported on screening for 'vertebral compression fracture' (Patrick 1983; Reinus 1998), and one study reported on 'unspecified vertebral fracture' (Gibson 1992).

Three studies investigated 'trauma' as a potential red flag (compression fracture: two studies, unspecified fracture: one study) (Gibson 1992; Patrick 1983; Reinus 1998). For 'vertebral compression fracture' (Patrick 1983; Reinus 1998) and 'unspecified vertebral fracture' (Gibson 1992), the diagnostic accuracy values varied greatly across the three studies and, therefore, meta-analysis was not performed. The sensitivity for the index test 'trauma' varied greatly from 0.07 (95% CI 0.02 to 0.18) to 1.00 (95% CI 0.59 to 1.00) and the specificity ranged from 0.51 (95% CI 0.41 to 0.62) to 0.60 (95% CI 0.56 to 0.65). The +LRs were also inconsistent ranging from 0.18 (95% CI 0.07 to 0.48) to 1.93 (95% CI 1.48 to 2.52).

One study investigated a range of index tests including components of subjective history and physical examination as potential red flags for 'vertebral compression fracture' (Patrick 1983). This study demonstrated one index test with some diagnostic value. The 'presence of contusion or abrasion' in an emergency department setting demonstrated high sensitivity (0.85, 95% CI 0.70 to 0.94) and specificity (0.97, 95% CI 0.95 to 0.98) indicating some diagnostic value. The +LR also demonstrated high diagnostic value (31.09, 95% CI 18.25 to 52.96). However, these results should be interpreted with caution as they were from a single study.

DISCUSSION

Summary of main results

This review is an update of a previous Cochrane Review of red flags to screen for vertebral fracture in people with low back pain (Williams 2013). Index tests included components of history taking or physical examination and included studies were in primary, secondary, and tertiary healthcare settings.

Our review only included six additional studies (total 14) investigating red flags to screen for vertebral fracture compared to the previous review. Meta-analysis of studies was not possible due to heterogeneity of the data and results from single studies suggested only a few of the red flags investigated may be informative. In the primary healthcare setting, results from single studies suggest 'trauma' demonstrated informative +LRs for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture'. Results from single studies suggest 'older age' demonstrated informative +LRs for studies in primary care for 'unspecified vertebral fracture'. Results from single studies suggest 'corticosteroid use' may be an informative red flag in primary care for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture'; however, diagnostic values varied, and CIs were imprecise. Results from a single study suggest red flags as part of a combination of index tests such as 'older age and female gender' in primary care demonstrated informative +LRs for 'unspecified vertebral fracture'. In the secondary healthcare setting, results from a single study suggest 'trauma' and 'older age' demonstrated informative +LRs for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture', respectively. Results from a single study suggest red flags as part of a combination of index tests such as 'older age and trauma' in secondary care demonstrated informative +LRs for 'unspecified vertebral fracture'. When four of five tests were positive in secondary care, they demonstrated informative +LRs for 'osteoporotic vertebral fracture'. In the tertiary care setting, results from a single study suggest 'presence of contusion/abrasion' was informative for 'vertebral compression fracture'.

Studies included in our review presented data on the current red flags recommended in guidelines to screen for vertebral fracture (e.g. osteoporosis, history of trauma, corticosteroid use, older age, and female gender) (Verhagen 2016). However, only one study presented data on osteoporosis despite its recommendation in current guidelines. Trauma, older age, corticosteroid use, and presence of contusion/abrasion were suggestive of informative +LRs in our study; however, CIs for corticosteroid use varied greatly. Results for female gender did not demonstrate any informative diagnostic values in our study. Current guidelines do not report on any specific red flags for vertebral stress fracture and only two studies in our review presented data on this type of fracture (Sugiura 2021; Therriault 2020).

A combination of red flags was more informative than red flags used in isolation. For example, when 'older age' (i.e. age greater than 74 years) and 'female gender' were combined the +LR increased from 9.39 (95% CI 2.69 to 32.75) to 16.17 (95% CI 4.47 to 58.43). When 'older age' (i.e. age greater than 70) and 'trauma' were combined, the +LR increased from 1.55 (95% CI 1.36 to 1.76) to 4.35 (95% CI 2.92 to 6.48). Another combination of index tests (age greater than 52 years, no presence of leg pain, BMI less than 23, does not exercise regularly, and female gender) found large +LRs if four of the five tests were positive (9.62, 95% CI 5.88 to 15.73). However, these findings should be interpreted with caution as they were based on single studies and there was a noticeable lack of available data in this field. Of the available data in our review, often +LR had wide CIs and imprecise estimates. Uncertainty in the risk of bias across the included studies should also be considered when interpreting results.

Many red flags used to screen for vertebral fracture have high false-positive rates and are currently based on limited data. Given the prevalence of serious pathology in primary care (1% to 5%) (Downie

2014; Hartvigsen 2018; Henschke 2009; Williams 2013), and in emergency departments (2.5% to 5.1%) (Galliker 2019) is quite low, the limited evidence makes it challenging for clinicians to identify vertebral fractures in people with low back pain. Clinicians also face the challenges of considering the consequences of unnecessary imaging and its associated impact on patient outcomes and patient costs.

Population and healthcare setting

The aim of our review was to identify red flags to assist clinicians to screen for vertebral fracture in people with low back pain. Although a few studies reported on the same or similar index test, many studies provided data based on different healthcare settings, different vertebral fracture types, or demonstrated large variability in diagnostic values, making meta-analysis of included data inappropriate. We included studies from different healthcare settings; however, based on the results of our review, most red flags recommended to screen for vertebral fracture as a cause of low back pain are not well-supported in any healthcare setting due to the lack of studies in this field and because most red flags identified have poor diagnostic value or have only been evaluated in single studies.

There was no clear pattern of differences between the studies conducted in primary and secondary care in the patient population and type of vertebral fracture investigated. The primary and secondary care studies included a mix of patient populations, with some studies only including older participants (e.g. Enthoven 2016; Jin 2020; Kilic 2021), and other studies only including younger populations (e.g. Sugiura 2021; Therriault 2020). Studies in primary and secondary care also investigated a variety of vertebral fractures such as 'osteoporotic vertebral fracture', 'stress fracture', and 'unspecified fracture'. Of the three studies conducted in tertiary care, two studies investigated 'vertebral compression fracture' and the patient populations were of similar age ranges across the three studies. Five of the 14 studies did not specify the type of vertebral fracture.

In the primary healthcare setting, the risk of bias of the included studies was not consistent in this domain (i.e. some studies had a risk of bias that was low, high, or unclear); however, the risk of bias was consistently unclear across most studies in the secondary and tertiary healthcare setting. It was evident that in four studies verification had occurred as in these studies there was a failure to compare the index tests and reference standards in all participants. This may have overestimated the accuracy of the index tests investigated.

Reference standard

Most studies used an adequate reference standard according to our methodological quality criteria (Appendix 4), except one study (Henschke 2009). This study demonstrated high concerns of applicability as clinical follow-up of participants with suspected vertebral fracture was used as the reference standard. It is possible that some participants with vertebral fracture in this study may have been missed, which would affect the estimates of diagnostic accuracy. Another study used a combination of different types of imaging methods (e.g. X-ray, MRI, CT, or a combination of these) based on the clinician's clinical preference as the reference standard (Therriault 2020). Although all participants received some form of imaging, not all participants received the same (form of imaging) reference standard.

Most studies used X-ray as the reference standard, however one of the main limitations of X-rays is that they are often unable to identify the acuity of a vertebral fracture (i.e. whether a fracture is acute or non-acute), whilst MRI has the ability to detect bone marrow oedema and more accurately distinguish acute versus chronic fractures (Ang 2016; Astur 2015). However, MRI has its limitations due to contraindications (e.g. indwelling electrical devices), feasibility, access, and cost (Kim 2022). CT may be used for detection of vertebral fractures, but often fails to determine the acuity or chronicity of fractures because of variability in fracture morphology, inhomogeneous background bone marrow attenuation, and different degrees of bone marrow oedema (Kim 2022). Another issue is that vertebral fractures such as compression fractures may often be incidental findings on CT, with some people being either asymptomatic or having symptoms not severe enough to seek medical care (Urrutia 2019).

Most studies also did not report on the method of imaging (e.g. X-ray views used, experience of the radiographer, thresholds for identifying vertebral fractures, sequences of imaging). Therefore, the influence of different imaging methods could not be investigated. All reference standards (e.g. X-ray, MRI, CT scan) available would likely be similar across all healthcare settings as these tests could be ordered in primary, secondary, or tertiary care. Given that all the reference standards are available across all healthcare settings, this would have little to no impact on concerns for applicability (i.e. there is low concern for applicability).

Index tests

If red flags have sufficiently high diagnostic accuracy their presence would indicate that further investigation is necessary to confirm serious pathology. If red flags are absent, then this should reduce the clinician's suspicion of serious pathology and negate the need for patients to undergo further investigation. Unfortunately, the accuracy of most red flags is low or uncertain (or both) so this use of red flags to guide diagnostic work-up is challenged.

The results of our review are based on a very small number of studies and currently do not support most of the red flags investigated. The sensitivity of most index tests was generally low, indicating that when vertebral fracture is present, there is a high number of people with vertebral fracture who will also go undiagnosed. Most tests also had small-to-modest +LRs and only a few index tests would increase the suspicion of having a vertebral fracture above the reported prevalence. In addition, the specificity of some of the informative index tests (e.g. older age, corticosteroid use, trauma, contusion/abrasion, a combination of some index tests) was high suggesting that if these tests are positive, it is likely that a fracture is present. However, the −LRs of most index tests in our review were not informative and are unlikely to be useful in clinical practice. Given the low prevalence of vertebral fracture, the consequences of being unable to rule out vertebral fractures using red flags may not be as detrimental as it is unlikely a negative test will meaningfully lower the probability of fracture any further or shift the provability of vertebral fracture.

The results of our review suggest some index tests were informative; however, these results were based on single studies. Trauma, older age, and corticosteroid use were suggestive of informative +LRs; however, the estimates were quite imprecise (e.g. +LRs 48.50, 95% CI 11.46 to 204.98 for corticosteroid use) and so should be interpreted with caution. The imprecision is

likely due to the low number of vertebral fractures identified in people presenting with low back pain. Recently, The International Federation of Orthopaedic Manipulative Physical Therapists (IFOMPT) developed an international framework for red flags for potential serious spinal pathologies (Finucane 2020). The recommendations of this framework were in line with the results of our review that trauma, older age, and corticosteroid use should increase clinical suspicion of vertebral fracture. The framework also recommended that history of osteoporosis, female sex, and previous spinal fracture should increase suspicion of vertebral fracture. In our review, history of osteoporosis demonstrated some informativeness (+LR 3.13, 95% CI 1.90 to 5.15) (Enthoven 2016); however, we did not find evidence to support the use of female sex and no studies in our review reported on previous spinal fracture. Clinicians may have to rely on frameworks such as the guidelines developed by IFOMPT as a guide for screening for vertebral fractures as there are currently very little primary data in this field. This creates challenges for researchers and clinicians to utilise red flags confidently in clinical practice.

Strengths and weaknesses of the review

A strength of this review is that we used the same search strategy used by Williams 2013, which was a sensitive and systematic search performed to include all possible studies. Our search identified previous systematic reviews on red flags for people presenting with low back pain (Downie 2014; Galliker 2019), and our search identified all relevant studies in those reviews.

A limitation of our review is the heterogeneous nature of the included studies and the paucity of diagnostic research evaluating red flags to screen for vertebral fracture. There were many index tests investigated by single studies, but few studies investigated the same index tests. When multiple studies did investigate the same index tests there was large variability in the diagnostic values between studies, which limited our ability to perform meta-analysis of diagnostic values. There was also little uniform agreement on the definition of the index tests, which may explain the diagnostic accuracy of such tests. For example, dose and duration of corticosteroid use is often not reported, yet this may influence the risk of certain vertebral fractures (e.g. osteoporotic vertebral fracture) (Vestergaard 2008). The acuity of osteoporotic vertebral compression fracture needs to be considered as prevalence estimates may differ in those with acute versus non-acute osteoporotic compression fracture. The studies investigating osteoporotic compression fractures did not report acuity and identifying the acuity of this type of fracture is often difficult due to factors such as the imaging modality used, scoring method used, and population investigated (Lentle 2019; Oei 2013). Many single studies investigated the same index tests across different healthcare settings, used different study designs, or presented diagnostic values for different types of vertebral fracture, which meant performing meta-analysis for diagnostic accuracy was not appropriate. Due to the limitations of the available evidence in our review, it was not possible to investigate the diagnostic accuracy of index tests or investigate potential sources of heterogeneity on the diagnostic accuracy of potential red flags. Some studies also included a highly selective sample and used an index test as part of both their inclusion criteria and as a red flag. For example, Enthoven 2016 only included adults over the age of 55 years and investigated trauma as an index test, therefore the index test investigated would be more akin to age over 55 years and 'trauma', not trauma alone. Studies that included a highly selective sample inflate prevalence estimates and this likely explains the high prevalence of vertebral fracture in those studies and may also be further influenced by the healthcare setting the study is performed in and the different types of vertebral fracture.

Two review authors independently performed the risk of bias assessment for this review using clear guidelines set a priori to minimise differences in scoring (Appendix 4). Most studies did not provide enough detail or information to make a clear decision on bias and were therefore scored at unclear risk of bias (Figure 2). Future research in this field needs to focus on higher quality and reporting in areas such as clarifying aspects of patient selection, standardising reference standards used across studies, and clarifying the interval time between the index test performed and the reference standard.

Applicability of findings to the review question

The current literature recommends that routine imaging for low back pain in clinical practice is not appropriate and should only be performed in patients where there is suspicion of serious pathology (Foster 2018). Unfortunately, due to the paucity of new studies and primary data in index tests to screen for vertebral fracture, it remains unclear which red flags are useful when screening for vertebral fracture in people presenting with low back pain. The limited research in this field may be a factor in the current high use of imaging in primary care. In our review, there was a total of 130 individual red flags that were investigated across 14 studies with very few studies investigating the same index test. Given the large number of index tests available to clinicians, it is understandable that many clinicians may refer for further imaging or investigations. Inappropriate or excessive imaging leads to excessive costs, potentially unnecessary radiation exposure, and can also lead to negative downstream consequences (i.e. more invasive procedures, poorer outcomes, etc.) (Jacobs 2020). However, there are consequences to also not using red flags in clinical practice with the potential for misdiagnosis of vertebral fracture, which is further complicated by the low prevalence of vertebral fracture in primary care. Without robust evidence clinicians will continue to face the challenge of being selective when referring patients for further investigations. The results from our review may guide clinicians when screening for vertebral fracture, as we found trauma, older age, corticosteroid use, history of osteoporosis, or a combination of red flags demonstrated some diagnostic value.

The results from our review support the view that referral for further investigations in people with suspected vertebral fracture should be based on a select number of red flags and using certain combinations of red flags may provide more precise diagnostic values. Given the overall quality of the current available evidence and that no meta-analysis of studies was possible, the recommendations of this review should be interpreted with caution. Although results from single studies do demonstrate some red flags may be useful and have the potential to increase clinical suspicion of vertebral fracture in people presenting with low back pain, more robust research is needed.

AUTHORS' CONCLUSIONS

Implications for practice

The available evidence suggests that only a few red flags are potentially useful in guiding clinical decisions regarding the need to further investigate people suspected to have a vertebral fracture. Most red flags are not useful in increasing the likelihood of vertebral fracture when present, nor are they useful in decreasing the likelihood of vertebral fracture when absent. Our review found a few red flags demonstrated some promise. In primary care, 'older age' was informative for 'unspecified vertebral fracture' and 'trauma' and 'corticosteroid use' were both informative for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture'. In secondary care, 'older age' was informative for 'osteoporotic vertebral fracture' and 'trauma' was informative for 'unspecified vertebral fracture'. In tertiary care, 'presence of contusion/abrasion' was informative for 'vertebral compression fracture'. Combinations of red flags were also informative and may be more useful in assisting clinicians with decisions than individual tests alone. Given the limitations of our review, it is still challenging to provide clear guidance as to which red flags should be used routinely in clinical practice. Further research with studies published in peer-reviewed journals is needed to improve and consolidate our current recommendations for screening for vertebral fractures to guide clinical care.

Implications for research

It is clear there is a paucity of primary diagnostic research in this field as we only found six additional studies since the previous version of the review (Williams 2013). Future research should consider which index tests may be clinically relevant based on the different types of vertebral fracture. There is very little consistency across studies in the index tests investigated and most studies do not provide clear details or thresholds for a positive or negative test. Most red flags investigated in the literature are individual tests; however, red flags (like many other tests) are rarely applied individually in clinical practice, yet these tests continue to be investigated in the research this way. Future studies should consider investigating combinations of index tests that seem most clinically appropriate or relevant for the different types of vertebral fracture. The results from our review highlight that a few promising index tests may be relevant in clinical practice, hence future studies should investigate these index tests to validate their use in a clinical care. Given the prevalence of vertebral fracture presenting as low back pain is low, primary diagnostic research requires large sample sizes to produce robust diagnostic accuracy estimates. These types of studies are challenging to conduct; however, they have the potential to reduce unnecessary imaging and its associated costs in a musculoskeletal condition (i.e. low back pain) that has been challenged with overimaging and its associated downstream consequences (Jacobs 2020).

ACKNOWLEDGEMENTS

We would like to thank Prof Petra Macaskill and Prof Maurits van Tulder for their assistance in the development of this protocol. We acknowledge peer reviewers: Prof Roger Chou, MD, Oregon Health & Science University.

We would like to thank the Cochrane Diagnostic Test Accuracy Group for reviewing this manuscript and providing feedback.

We would like to thank Anne Lawson (Cochrane Central Production Service) for copy editing and proofreading this review.

REFERENCES

References to studies included in this review

Deyo 1986 *{published data only}*

Deyo RA, Diehl AK. Lumbar spine films in primary care: current use and effects of selective ordering criteria. *Journal of General Internal Medicine* 1986;**1**:20-5.

Enthoven 2016 *{published data only}*

Enthoven WT, Geuze J, Scheele J, Bierma-Zeinstra SM, Bueving HJ, Bohnen AM, et al. Prevalence and "red flags" regarding specified causes of back pain in older adults presenting in general practice. *Physical Therapy* 2016;**96**(3):305-12.

Gibson 1992 *{published data only}*

Gibson M, Zoltie N. Radiography for back pain presenting to accident and emergency departments. *Archives of Emergency Medicine* 1992;**9**(1):28-31.

Henschke 2009 *{published data only}*

Henschke N, Maher CG, Refshauge KM, Herbert RD, Cumming RG, Bleasel J, et al. Prevalence of and screening for serious spinal pathology in patients presenting to primary care settings with acute low back pain. *Arthritis & Rheumatism* 2009;**60**(10):3072-80. [MEDLINE: 0004-3591]

Jin 2020 *{published data only}*

Jin H, Ma X, Liu Y, Yin X, Zhu J, Wang Z, et al. Back pain-inducing test, a novel and sensitive screening test for painful osteoporotic vertebral fractures: a prospective clinical study. *Journal of Bone and Mineral Research* 2020;**35**(3):488-97.

Kilic 2021 *{published data only}*

Kilic Z, Alkan BM. The frequency of spontaneous vertebral fracture in geriatric patients and the relationship of vertebral fractures with age: a retrospective study. *Turkish Journal of Osteoporosis* 2021;**27**:90-5.

Patrick 1983 *{published data only}*

Patrick JD, Doris PE, Mills ML, Friedman J, Johnston C. Lumbar spine x-rays: a multihospital study. *Annals of Emergency Medicine* 1983;**12**(2):84-7.

Premkumar 2018 *{published data only}*

Premkumar A, Godfrey W, Gottschalk MB, Boden SD. Red flags for low back pain are not always really red: a prospective evaluation of the clinical utility of commonly used screening questions for low back pain. *Journal of Bone and Joint Surgery* 2018;**100**(5):368-74.

Reinus 1998 *{published data only}*

Reinus WR, Strome G, Zwemer Jr FL. Use of lumbosacral spine radiographs in a level II emergency department. *American Journal of Roentgenology* 1998;**179**(2):443-7.

Roman 2010 *{published data only}*

Roman M, Brown C, Richardson W, Isaacs R, Howes C, Cook C. The development of a clinical decision making algorithm for detection of osteoporotic vertebral compression fracture or wedge deformity. *Journal of Manual & Manipulative Therapy* 2010;**18**(1):44-9.

Scavone 1981 *{published data only}*

Scavone JG, Latshaw RF, Rohrer GV. Use of lumbar spine films. Statistical evaluation at a university teaching hospital. *JAMA* 1981;**246**:1105-8.

Sugiura 2021 *{published data only}*

Sugiura S, Aoki Y, Toyooka T, Shiga T, Oyama T, Ishizaki T, et al. Utility of physical examination findings for predicting low-back pain in adolescent patients with early-stage spondylolysis: a retrospective comparative cohort study. *Spine Surgery and Related Research* 2021;**5**(6):412-7.

Therriault 2020 *{published data only}*

Therriault T, Rospert A, Selhorst M, Fischer A. Development of a preliminary multivariable diagnostic prediction model for identifying active spondylolysis in young athletes with low back pain. *Physical Therapy in Sport* 2020;**45**:1-6.

van den Bosch 2004 *{published data only}*

van den Bosch MA, Hollingworth W, Kinmonth AL, Dixon AK. Evidence against the use of lumbar spine radiography for low back pain. *Clinical Radiology* 2004;**59**(1):69-76.

References to studies excluded from this review

Althoff 2017 *{published data only}*

Althoff S, Overberger R, Sochor M, Bose D, Werner J. GLASS clinical decision rule applied to thoracolumbar spinal fractures in patients involved in motor vehicle crashes. *Western Journal of Emergency Medicine: Integrating Emergency Care with Population Health* 2017;**18**(6):1108-13.

Baleanu 2020 *{published data only}*

Baleanu F, Moreau M, Kinnard V, Iconaru L, Karmali R, Paesmans M, et al. What is the validity of self-reported fractures? *Bone Reports* 2020;**12**(1):100256.

Bian 2022 *{published data only}*

Bian FC, Bian GY, An YS, Wang DY, Fang JH. Establishment and validation of a nomogram for the risk of new vertebral compression fractures after percutaneous vertebroplasty in patients with osteoporotic vertebral compression fractures: a retrospective study. *Geriatric Orthopaedic Surgery & Rehabilitation* 2022;**13**:1-11.

Borgen 2020 *{published data only}*

Borgen TT, Bjørnerem Å, Solberg LB, Andreasen C, Brunborg C, Stenbro MB, et al. Determinants of trabecular bone score and prevalent vertebral fractures in women with fragility fractures: a cross-sectional sub-study of NoFRACT. *Osteoporosis International* 2020;**31**(3):505-14.

Bottai 2016 *{published data only}*

Bottai V, Giannotti S, Raffaeta G, Mazzantini M, Casella F, Paola GD, et al. Underdiagnosis of osteoporotic vertebral

fractures in patients with fragility fractures: retrospective analysis of over 300 patients. *Clinical Cases in Mineral and Bone Metabolism* 2016;**13**(2):119-22.

Cheung 2018 *{published data only}*

Cheung KK, Dhawan RT, Wilson LF, Peirce NS, Rajeswaran G. Pars interarticularis injury in elite athletes – the role of imaging in diagnosis and management. *European Journal of Radiology* 2018;**108**:28-42.

Clark 2017 *{published data only}*

Clark EM, Cummings SR, Schousboe JT. Spinal radiographs in those with back pain-when are they appropriate to diagnose vertebral fractures? *Osteoporosis international* 2017;**28**(8):2293-7.

Della-Giustina 2013 *{published data only}*

Della-Giustina D. Acute low back pain: recognizing the "red flags" in the workup. *Consultant 360* 2013;**53**(6):436-40.

Della-Giustina 2015 *{published data only}*

Della-Giustina D. Evaluation and treatment of acute back pain in the emergency department. *Emergency Medicine Clinics of North America* 2015;**33**(2):311-26.

Durham 1995 *{published data only}*

Durham RM, Luchtefeld WB, Wibbenmeyer L, Maxwell P, Shapiro MJ, Mazuski JE. Evaluation of the thoracic and lumbar spine after blunt trauma. *American Journal of Surgery* 1995;**170**(6):681-4. [MEDLINE: 0022-3085]

Enoki 2022 *{published data only}*

Enoki S, Kuramochi R, Nakajyuku S, Mitsuyama H. The prevalence of spondylolysis and intervertebral disc degeneration in male pole vaulters. *Journal of Back & Musculoskeletal Rehabilitation* 2022;**35**(1):147-51.

Finucane 2020 *{published data only}*

Finucane LM, Downie A, Mercer C, Greenhalgh SM, Boissonnault WG, Pool-Goudzwaard AL, et al. International framework for red flags for potential serious spinal pathologies. *Journal of Orthopaedic and Sports Physical Therapy* 2020;**50**(7):350-72.

Frankel 1994 *{published data only}*

Frankel HL, Rozycki GS, Ochsner MG, Harviel JD, Champion HR. Indications for obtaining surveillance thoracic and lumbar spine radiographs. *Journal of Trauma* 1994;**37**(4):673-6. [MEDLINE: 0022-5282]

Gaddikeri 2018 *{published data only}*

Gaddikeri S, Matesan M, Alvarez J, Hippe DS, Vesselle HJ. MDP-SPECT versus hybrid MDP-SPECT/CT in the evaluation of suspected pars interarticularis fracture in young athletes. *Journal of Neuroimaging* 2018;**28**(6):635-9.

Galliker 2019 *{published data only}*

Galliker G, Scherer DE, Trippolini MA, Rasmussen-Barr E, LoMartire R, Wertli MM. Low back pain in the emergency department: prevalence of serious spinal pathologies and diagnostic accuracy of red flags. *American Journal of Medicine* 2019;**133**(1):60-72.

Gazzola 2015 *{published data only}*

Gazzola L, Savoldi A, Bai F, Magenta A, Dziubak M, Pietrogrande L, et al. Assessment of radiological vertebral fractures in HIV-infected patients: clinical implications and predictive factors. *HIV Medicine* 2015;**16**(9):563-71.

Gestring 2002 *{published data only}*

Gestring ML, Gracias VH, Feliciano MA, Reilly PM, Shapiro MB, Johnson JW, et al. Evaluation of the lower spine after blunt trauma using abdominal computed tomographic scanning supplemented with lateral scanograms. *Journal of Trauma* 2002;**53**(1):9-14.

Gill 2013 *{published data only}*

Gill DS, Mitra B, Reeves F, Cameron PA, Fitzgerald M, Liew S, et al. Can initial clinical assessment exclude thoracolumbar vertebral injury? *Emergency Medicine Journal* 2013;**30**(8):679-82.

Hiett 2014 *{published data only}*

Hiett J, Chan D, Hourigan P. Back pain in the under 20s would appear to be a 'red flag' of dubious significance: results of a retrospective case note analysis. *International Musculoskeletal Medicine* 2014;**36**(1):19-22.

Holmes 2003 *{published data only}*

Holmes JF, Panacek EA, Miller PQ, Lapidis AD, Mower WR. Prospective evaluation of criteria for obtaining thoracolumbar radiographs in trauma patients. *Journal of Emergency Medicine* 2003;**24**(1):1-7.

Hsu 2003 *{published data only}*

Hsu JM, Joseph T, Ellis AM. Thoracolumbar fracture in blunt trauma patients: guidelines for diagnosis and imaging. *Injury* 2003;**34**(6):426-33.

Johannesdottir 2021 *{published data only}*

Johannesdottir F, Allaire B, Kopperdahl DL, Keaveny TM, Sigurdsson S, Bredella MA, et al. Bone density and strength from thoracic and lumbar CT scans both predict incident vertebral fractures independently of fracture location. *Osteoporosis International* 2021;**32**(2):261-9.

Johansson 2018 *{published data only}*

Johansson L, Sundh D, Nilsson M, Mellström D, Lorentzon M. Vertebral fractures and their association with health-related quality of life, back pain and physical function in older women. *Osteoporosis International* 2018;**29**(1):89-99.

Kendler 2016 *{published data only}*

Kendler DL, Bauer DC, Davison KS, Dian L, Hanley DA, Harris ST, et al. Vertebral fractures: clinical importance and management. *American Journal of Medicine* 2016;**129**(2):221.e1-10.

Khera 2022 *{published data only}*

Khera TK, Hunt LP, Davis S, Gooberman-Hill R, Thom H, Xu Y, et al. A clinical tool to identify older women with back pain at high risk of osteoporotic vertebral fractures (Vfrac): a population-

based cohort study with exploratory economic evaluation. *Age and Ageing* 2022;**1**(51):1-8.

Kherad 2015 *{published data only}*

Kherad M, Rosengren BE, Hasserius R, Nilsson JA, Redlund-Johnell I, Ohlsson C, et al. Low clinical relevance of a prevalent vertebral fracture in elderly men-the MrOs Sweden study. *Spine Journal* 2015;**15**(2):281-9.

Korner 2014 *{published data only}*

Korner H. Red flags in the diagnostics of back pain. *Die Radiologie* 2014;**54**(11):1078-81.

Kuru 2014 *{published data only}*

Kuru P, Akyüz G, Cerşit HP, Çelenlioğlu AE, Cumhur A, Biricik Ş, et al. Fracture history in osteoporosis: risk factors and its effect on quality of life. *Balkan Medical Journal* 2014;**31**(4):295-301.

Manji 2022 *{published data only}*

Manji R, Ponzano M, Ashe MC, Wark JD, Kendler D, Papaioannou A, et al. Exploring the association between pain and fracture characteristics in women with osteoporotic vertebral fractures. *Physiotherapy Canada Spring* 2022;**74**(2):165-72.

Mazziotti 2016 *{published data only}*

Mazziotti G, Dordoni C, Doga M, Galderisi F, Venturini M, Calzavara-Pinton P, et al. High prevalence of radiological vertebral fractures in adult patients with Ehlers-Danlos syndrome. *Bone* 2016;**84**:88-92.

Min 2017 *{published data only}*

Min A, Chan VW, Aristizabal R, Peramaki ER, Agulnik DB, Strydom N, et al. Clinical decision support decreases volume of imaging for low back pain in an urban emergency department. *Journal of the American College of Radiology* 2017;**14**(7):889-99.

Mizukami 2014 *{published data only}*

Mizukami S, Abe Y, Tsujimoto R, Arima K, Kanagae M, Chiba G, et al. Accuracy of spinal curvature assessed by a computer-assisted device and anthropometric indicators in discriminating vertebral fractures among individuals with back pain. *Osteoporosis International* 2014;**25**(6):1727-34.

Moretti 2019 *{published data only}*

Moretti A, de Sire A, Curci C, Toro G, Gimigliano F, Lolascon G. Effectiveness of denosumab on back pain-related disability and quality-of-life in patients with vertebral fragility fractures. *Current Medical Research and Opinion* 2019;**35**(1):151-5.

Padlina 2017 *{published data only}*

Padlina I, Gonzalez-Rodriguez E, Hans D, Metzger M, Stoll D, Aubry-Rozier B, et al. The lumbar spine age-related degenerative disease influences the BMD not the TBS: the Osteolaus cohort. *Osteoporosis International* 2017;**28**(3):909-15.

Pressney 2014 *{published data only}*

Pressney I, Hargunani R, Khoo M, Tyler PA. Radiological imaging in the spine. *Orthopaedics and Trauma* 2014;**28**(2):106-15.

Rajakulasingam 2022 *{published data only}*

Rajakulasingam R, Nightingale P, Bhatt N, Choudhary S. Cortical step sign in spinal clearance on trauma computed tomography – indicator of acute thoracolumbar compression fracture. *Trauma* 2022;**25**(2):146040862110566.

Rapala 2012 *{published data only}*

Rapala K, Walczak P, Truszczynska A, Lukawski S, Nowak-Misiak M. Diagnostic and therapeutic problems of back pain syndromes and their distribution according to a colour coding system of flags. *Ortopedia Traumatologia Rehabilitacja* 2012;**14**(3):215-27.

Reito 2018 *{published data only}*

Reito A, Kyrola K, Pekkanen L, Paloneva J. Specific spinal pathologies in adult patients with an acute or subacute atraumatic low back pain in the emergency department. *International Orthopaedics* 2018;**42**(12):2843-9.

Rudman 2019 *{published data only}*

Rudman HA, Treadgold L, Birrell F, Tuck SP, Hind K, Pearce MS, et al. Obesity, bone density relative to body weight and prevalent vertebral fracture at age 62 years: the Newcastle thousand families study. *Osteoporosis International* 2019;**30**(4):829-36.

Sakai 2016 *{published data only}*

Sakai T, Goda Y, Tezuka, F, Abe M, Yamashita K, Takata Y, et al. Clinical features of patients with pars defects identified in adulthood. *European Journal of Orthopaedic Surgery & Traumatology* 2016;**26**(3):259-62.

Samuels 1993 *{published data only}*

Samuels LE. Routine radiologic evaluation of the thoracolumbar spine in blunt trauma patients: a reappraisal. *Journal of Trauma* 1993;**34**(1):85-9. [MEDLINE: 0362-2436]

Scaturro 2020 *{published data only}*

Scaturro D, Lauricella L, Tumminelli LG, Tomasello S, Mauro GL. Is there a relationship between mild-moderate back pain and fragility fractures? Original investigation. *Acta Medica Mediterranea* 2020;**36**(3):2149-53.

Shah 2021 *{published data only}*

Shah A, Wu F, Jones G, Cicuttini F, Toh LS, Laslett LL. The association between incident vertebral deformities, health-related quality of life and functional impairment: a 10.7-year cohort study. *Osteoporosis International* 2021;**32**(11):2247-55.

Soto-Subiabre 2020 *{published data only}*

Soto-Subiabre M, Mayoral V, Fiter J, Valencia L, Subirana I, Gómez-Vaquero C. Vertebral fracture: clinical presentation and severity are linked to fracture risk factors. *Osteoporosis International* 2020;**31**(9):1759-68.

St Jeor 2020 *{published data only}*

St Jeor JD, Jackson TJ, Xiong AE, Kadri A, Freedman BA, Sebastian AS, et al. Osteoporosis in spine surgery patients: what is the best way to diagnose osteoporosis in this population? *Neurosurgical Focus* 2020;**49**(2):E4.

Sucato 2012 *{published data only}*

Sucato DJ, Micheli LJ, Estes AR, Tolo VT. Spine problems in young athletes. *Instructional Course Lectures* 2012;**61**:499-511.

Terakado 2017 *{published data only}*

Terakado A, Orita S, Inage K, Kubota G, Kanzaki T, Mori H, et al. A clinical prospective observational cohort study on the prevalence and primary diagnostic accuracy of occult vertebral fractures in aged women with acute lower back pain using magnetic resonance imaging. *Pain Research & Management* 2017;**2017**:9265259.

Terregino 1995 *{published data only}*

Terregino CA, Ross SE, Lipinski MF, Foreman J, Hughes R. Selective indications for thoracic and lumbar radiography in blunt trauma. *Annals of Emergency Medicine* 1995;**26**(2):126-9.

Trout 2015 *{published data only}*

Trout AT, Sharp SE, Anton CG, Gelfand MJ, Mehlman CT. Spondylolysis and beyond: value of SPECT/CT in evaluation of low back pain in children and young adults. *Radiographics* 2015;**35**(3):819-34.

Upadhye 2016 *{published data only}*

Upadhye S, Kumbhare D. What are the most useful red flags for suspected vertebral fracture in patients with low back pain in the emergency department? *Annals of Emergency Medicine* 2016;**67**(1):81-2.

van der Jagt-Willems 2012 *{published data only}*

van der Jagt-Willems HC, van Hengel M, Vis M, van Munster BC, van Campen JP, Tulner LR, et al. Why do geriatric outpatients have so many moderate and severe vertebral fractures? Exploring prevalence and risk factors. *Age & Ageing* 2012;**41**(2):200-6.

Yoshimura 2017 *{published data only}*

Yoshimura N, Muraki S, Oka H, Iidaka T, Kodama R, Kawaguchi H, et al. Is osteoporosis a predictor for future sarcopenia or vice versa? Four-year observations between the second and third ROAD study surveys. *Osteoporosis International* 2017;**28**(1):189-99.

Zong 2016 *{published data only}*

Zong Y, Tang Y, Xue Y, Ding H, Li Z, He D, et al. Depression is associated with increased incidence of osteoporotic thoracolumbar fracture in postmenopausal women: a prospective study. *European Spine Journal* 2016;**25**(11):3418-23.

Zorin 2020 *{published data only}*

Zorin VI, Mushkin AY, Novitskaya TA. Pathological fractures of the spine in children: (review of the literature and clinical and morphological monocenter cohort analysis). *Pediatric Traumatology, Orthopaedics and Reconstructive Surgery* 2020;**8**(1):5-14.

Additional references

Alqarni 2015

Alqarni AM, Schneiders AG, Cook CE, Hendrick PA. Clinical tests to diagnose lumbar spondylolysis and spondylolisthesis: a systematic review. *Physical Therapy in Sport* 2015;**16**(3):268-75.

Ang 2016

Ang EC, Robertson AF, Malara FA, O'Shea T, Roebert JK, Schneider ME, et al. Diagnostic accuracy of 3-T magnetic resonance imaging with 3D T1 VIBE versus computer tomography in pars stress fracture of the lumbar spine. *Skeletal Radiology* 2016;**45**(11):1533-40.

Astur 2015

Astur DC, Zanatta F, Arliani GG, Moraes ER, Pochini Ade C, Ejnisman B. Stress fractures: definition, diagnosis and treatment. *Revista Brasileira de Ortopedia* 2015;**51**(1):3-10.

Bazzocchi 2012

Bazzocchi A, Spinnato P, Fuzzi F, Diano D, Morselli-Labate AM, Sassi C, et al. Vertebral fracture assessment by new dual-energy X-ray absorptiometry. *Bone* 2012;**50**(4):836-41.

Berrington de González 2009

Berrington de González A, Mahesh M, Kim KP, Bhargavan M, Lewis R, Mettler F, et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. *Archives of Internal Medicine* 2009;**169**(22):2071-77.

Buchbinder 2018

Buchbinder R, van Tulder M, Öberg B, Costa LM, Woolf A, Schoene M, et al. Low back pain: a call for action. *Lancet* 2018;**391**(10137):2384-8.

Cheung 2018

Cheung KK, Dhawan RT, Wilson LF, Peirce NS, Rajeswaran G. Pars interarticularis injury in elite athletes – the role of imaging in diagnosis and management. *European Journal of Radiology* 2018;**108**:28-42.

Chu 2006

Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006;**59**(12):1331-3.

Cook 2018

Cook CE, George SZ, Reiman MP. Red flag screening for low back pain: nothing to see here, move along: a narrative review. *British Journal of Sports Medicine* 2018;**52**(8):493-6.

Damiano 2006

Damiano J, Kolta S, Porcher R, Tournoux C, Dougados M, Roux C. Diagnosis of vertebral fractures by vertebral fracture assessment. *Journal of Clinical Densitometry* 2006;**9**(1):66-71.

de Schepper 2016

de Schepper EI, Koes BW, Veldhuizen EF, Oei EH, Bierma-Zeinstra SM, Luijsterburg PA. Prevalence of spinal pathology in patients presenting for lumbar MRI as referred from general practice. *Family Practice* 2016;**33**(1):51-6.

Deeks 2004

Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ* 2004;**329**(7458):168-9.

Deeks 2005

Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* 2005;**8**(9):882-93.

Deeks 2022

Deeks JJ, Bossuyt PM, Leeflang M, Takwoingi Y, Flemyng E. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. 2nd edition. London: Cochrane, 2022.

DePalma 2020

DePalma MG. Red flags of low back pain. *Journal of the American Academy of PAs* 2020;**33**(8):8-11. [DOI: 10.1097/01.JAA.0000684112.91641.4c]

Downie 2014

Downie A, Williams CM, Henschke N, Hancock MJ, Ostelo RW, de Vet HC, et al. Red flags to screen for malignancy and fracture in patients with low back pain: systematic review. *BMJ* 2014;**48**(20):1518-56.

Ferreira 2021

Ferreira GE, Herbert R, Machado GC, Richards B, Harris IA, O'Keeffe M, et al. Low back pain presentations to New South Wales emergency departments: trends over time and geographical variation. *Emergency Medicine Australasia* 2021;**33**(5):868-74.

Finucane 2020

Finucane LM, Downie A, Mercer C, Greenhalgh SM, Boissonnault WG, Pool-Goudzwaard AL, et al. International Framework for Red Flags for Potential Serious Spinal Pathologies. *Journal of Orthopaedic & Sports Physical Therapy* 2020;**50**(7):350-72.

Foster 2018

Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, et al, Lancet Low Back Pain Series Working Group. Prevention and treatment of low back pain: evidence, challenges, and promising directions. *Lancet* 2018;**9**(391):2368-83.

Fuerst 2009

Fuerst T, Wu C, Genant HK, von Ingersleben G, Chen Y, Johnston C, et al. Evaluation of vertebral fracture assessment by dual X-ray absorptiometry in a multicenter setting. *Osteoporosis International* 2009;**20**(7):1199-205.

Galliker 2019

Galliker G, Scherer DE, Trippolini MA, Rasmussen-Barr E, LoMartire R, Wertli MM. Low back pain in the emergency department: prevalence of serious spinal pathologies and diagnostic accuracy of red flags. *American Journal of Medicine* 2019;**133**(1):60-72.

Genant 1993

Genant HK, Wu CY, van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. *Journal of Bone and Mineral Research* 1993;**8**(9):1137-48.

Grunau 2017

Grunau GL, Darlow B, Flynn T, O'Sullivan K, O'Sullivan PB, Forster BB. Red flags or red herrings? Redefining the role of red flags in low back pain to reduce overimaging. *British Journal of Sports Medicine* 2017;**52**(8):488-9.

Harbord 2007

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

Hartvigsen 2018

Hartvigsen J, Hanock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, et al. What low back pain is and why we need to pay attention. *Lancet* 2018;**391**(10137):2356-67.

Henschke 2009

Henschke N, Maher CG, Refshauge KM, Herbert RD, Cumming RG, Bleasel J, et al. Prevalence of and screening for serious spinal pathology in patients presenting to primary care settings with acute low back pain. *Arthritis & Rheumatology* 2009;**60**(10):3072-80.

Jacobs 2020

Jacobs JC, Jarvik JG, Chou R, Boothroyd D, Lo J, Nevedal A, et al. Observational study of the downstream consequences of inappropriate MRI of the lumbar spine. *Journal of General Internal Medicine* 2020;**35**(12):3605-12.

Jaeschke 1994

Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *Journal of the American Medical Association* 1994;**271**(9):703-7.

Jarvik 2003

Jarvik JG, Hollingworth W, Martin B, Emerson SS, Gray DT, Overman S, et al. Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomised controlled trial. *Journal of the American Medical Association* 2003;**289**(21):2810-8.

Jiang 2004

Jiang G, Eastell R, Barrington NA, Ferrar L. Comparison of methods for the visual identification of prevalent vertebral fracture in osteoporosis. *Osteoporosis International* 2004;**15**(11):887-96.

Johansson 2020

Johansson L, Sundh D, Magnusson P, Rukmangatharajan K, Mellström D, Nilsson AG, et al. Grade 1 vertebral fractures identified by densitometric lateral spine imaging predict incident major osteoporotic fracture independently of clinical risk factors and bone mineral density in older women. *Journal of Bone and Mineral Research* 2020;**35**(10):1942-51.

Kim 2015

Kim DE, Kim HS, Kim SW, Hyun Sook Kim HS. Clinical analysis of acute radiculopathy after osteoporotic lumbar compression fracture. *Journal of Korean Neurosurgical Society* 2015;**57**(1):32-5.

Kim 2022

Kim AY, Yoon MA, Ham SJ, Cho YC, Ko Y, Park B, et al. Prediction of the acuity of vertebral compression fractures on CT using radiologic and radiomic features. *Academic Radiology* 2022;**29**(10):1512-20.

Lee 2016

Lee JH, Lee YK, Oh SH, Ahn J, Lee E, Pyo JH, et al. A systematic review of diagnostic accuracy of vertebral fracture assessment (VFA) in postmenopausal women and elderly men. *Osteoporosis International* 2016;**27**(5):1691-9.

Lems 2021

Lems WF, Paccou J, Zhang J, Fuggle NR, Chandran M, Harvey NC, et al. Vertebral fracture: epidemiology, impact and use of DXA vertebral fracture assessment in fracture liaison services. *Osteoporosis International* 2021;**32**(3):399-411.

Lentle 2019

Lentle B, Koromani F, Brown JP, Oei L, Ward L, Goltzman D, et al, Vertebral Fracture Research Groups of the CaMos, STOPP, and Rotterdam Studies. The radiology of osteoporotic vertebral fractures revisited. *Journal of Bone and Mineral Research* 2019;**34**(3):409-18.

Marongiu 2018

Marongiu G, Congia S, Verona M, Lombardo M, Podda D, Capone A. The impact of magnetic resonance imaging in the diagnostic and classification process of osteoporotic vertebral fractures. *Injury* 2018;**49**(3):26-31.

McInnes 2018

McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, PRISMA-DTAG. Preferred Reporting Items for a Systematic Review and Meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA Statement. *Journal of the American Medical Association* 2018;**319**(4):388-96.

Murthy 2012

Murthy NS. Imaging of stress fractures of the spine. *Radiologic Clinics of North America* 2012;**50**(4):799-821.

Oei 2013

Oei L, Rivadeneira F, Ly F, Breda SJ, Zillikens MC, Hofman A, et al. Review of radiological scoring methods of osteoporotic vertebral fractures for clinical and research settings. *European Radiology* 2013;**23**(2):476-86.

Oliveira 2018

Oliveira CB, Maher CG, Pinto RZ, Traeger AC, Lin CW, Chenot JF, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *European Spine Journal* 2018;**27**(11):2791-803.

Parreira 2019

Parreira PC, Maher CG, Traeger AC, Hancock MJ, Downie A, Koes BW, et al. Evaluation of guideline-endorsed red flags to screen for fracture in patients presenting with low back pain. *British Journal of Sports Medicine* 2019;**53**(10):648-54.

Patel 2016

Patel ND, Broderick DF, Burns J, Deshmukh TK, Blair IB, Harvey B, et al. ACR appropriateness criteria low back pain. *Journal of the American College of Radiology* 2016;**13**(9):1069-78.

Premkumar 2018

Premkumar A, Godfrey W, Gottschalk MB, Boden SD. Red flags for low back pain are not always really red: a prospective evaluation of the clinical utility of commonly used screening questions for low back pain. *Journal of Bone and Joint Surgery* 2018;**100**(5):368-74.

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.

Review Manager 2014 [Computer program]

Review Manager 5 (RevMan 5). Version 5.4. Copenhagen: The Cochrane Collaboration, 2014.

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.

Schousboe 2006

Schousboe JT, Ensrud KE, Nyman JA, Kane RL, Melton LJ. Cost-effectiveness of vertebral fracture assessment to detect prevalent vertebral deformity and select postmenopausal women with a femoral neck T-score>-2.5 for alendronate therapy: a modelling study. *Journal of Clinical Densitometry* 2006;**9**(2):133-43.

Schousboe 2016

Schousboe JT. Epidemiology of Vertebral Fractures. *Journal of Clinical Densitometry* 2016;**19**(1):8-22.

Selhorst 2019

Selhorst M, Fischer A, MacDonald J. Prevalence of spondylolysis in symptomatic adolescent athletes: an assessment of sport risk in non-elite athletes. *Clinical Journal of Sport Medicine* 2019;**29**(5):421-5.

Underwood 2009

Underwood M. Diagnosing acute nonspecific low back pain: time to lower the red flags? *Arthritis & Rheumatology* 2009;**60**(10):2855-7.

Urrutia 2019

Urrutia J, Besa P, Piza C. Incidental identification of vertebral compression fractures in patients over 60 years old using computed tomography scans showing the entire thoraco-

lumbar spine. *Archives of Orthopaedic and Trauma Surgery* 2019;**139**(11):1497-503.

Verhagen 2016

Verhagen AP, Downie A, Popal N, Maher C, Koes BW. Red flags presented in current low back pain guidelines: a review. *European Spine Journal* 2016;**25**(9):2788-802.

Vestergaard 2008

Vestergaard P, Rejnmark L, Mosekilde L. Fracture risk associated with different types of oral corticosteroids and effect of termination of corticosteroids on the risk of fractures. *Calcified Tissue International* 2008;**82**(4):249-57.

Wall 2006

Wall BF, Kendall GM, Edwards AA, Bouffler S, Muirhead CR, Meara JR. What are the risks from medical X-rays and other low dose radiation? *British Journal of Radiology* 2006;**79**(940):285-94.

Warriner 2011

Warriner AH, Patkar NM, Curtis JR, Delzell E, Gary L, Kilgore M, et al. Which fractures are most attributable to osteoporosis? *Journal of Clinical Epidemiology* 2011;**64**(1):46-53.

Whiting 2011

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011;**155**(8):529-36.

Whiting 2013

Whiting PF, Rutjes AW, Westwood ME, Mallett S, QUADAS-2SG. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology* 2013;**66**(10):1093-104.

Wu 2020

Wu A, March L, Zheng X, Huang J, Wang X, Zhao J, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. *Annals of Translational Medicine* 2020;**8**(6):299.

References to other published versions of this review

Henschke 2010b

Henschke N, Williams CM, Maher CG, van Tulder MW, Koes BW, Macaskill P, et al. Red flags to screen for vertebral fracture in patients presenting with low-back pain. *Cochrane Database of Systematic Reviews* 2010, Issue 8. Art. No: CD008643. [DOI: 10.1002/14651858.CD008643]

Williams 2013

Williams CM, Henschke N, Maher CG, van Tulder MW, Koes BW, Macaskill P, et al. Red flags to screen for vertebral fracture in patients presenting with low-back pain. *Cochrane Database of Systematic Reviews* 2013, Issue 1. Art. No: CD008643. [DOI: 10.1002/14651858.CD008643.pub2]

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Deyo 1986

Study characteristics	
Patient Sampling	**Study design:** prospective cohort **Enrolment procedures:** consecutive sample **Sample size:** 621; 311 received imaging **Inclusion criteria:** people presenting with LBP **Exclusion criteria:** maximal pain above T12, evidence of urinary tract disease, women aged < 45 years who did not use contraception and whose last menstrual period occurred > 10 days previous, and participants of a concurrent study
Patient characteristics and setting	**Age group:** adolescents and adults (mean 40.5 years; range 15–86) years **Gender:** 293 male, 328 female **Duration of pain:** < 1 month: 447, ≥ 1 month: 146, unknown: 28 **History of LBP:** 289/621 **Healthcare setting:** walk-in university clinic

Deyo 1986 *(Continued)*

	Country: USA **Type of vertebral fracture:** vertebral fracture; type unclear from text **Reasons for withdrawal:** unclear from text
Index tests	**Index tests** • Significant trauma • Age > 50 years • Corticosteroid use **Methods of execution:** checklist **Threshold for positive result on index test:** presence or absence of red flags from checklist **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture or malignancy on: • X-ray – anteroposterior and lateral lumbar views
Flow and timing	**Timing of index test and reference test:** 84% of imaging was obtained on the day of the index visit or within 6 days thereafter.
Comparative	
Notes	**Prevalence of vertebral fracture:** 4.5%

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern

Deyo 1986 *(Continued)*

DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	No		
Could the patient flow have introduced bias?		High risk	

Enthoven 2016

Study characteristics	
Patient Sampling	**Study design:** prospective observational cohort **Enrolment procedures:** consecutive sample **Sample size:** 669 **Inclusion criteria:** people consulting a GP with a new episode of back pain **Exclusion criteria:** unable to fill out the questionnaires due to cognitive impairment, were unable to read and write in Dutch, unable to undergo physical examination (e.g. people using wheelchairs)
Patient characteristics and setting	**Age group:** adults aged > 55 years **Gender:** 269 men, 400 women **Duration of pain:** > 3 months: 154/669 **History of LBP:** 87/669 **Healthcare setting:** primary care: general practice **Country:** Netherlands **Type of vertebral fracture:** osteoporotic vertebral fracture (30/33); unclear from text type of fracture for 3/30 participants

Enthoven 2016 *(Continued)*

	Reasons for withdrawal: 6 (0.9%) participants were excluded from the analyses because they moved to another city during the 1-year follow-up and, due to a change of GP practice, or diagnosis could not be retrieved.
Index tests	**Index tests** • Aged ≥ 75 years • Female sex • Prolonged corticosteroid use • Trauma • Osteoporosis • Sudden decrease in height • Percussion tenderness of spine • Severe disability • Acute onset of pain • Back pain intensity score ≥ 7 • Osteoarthritis in hip/knee • Thoracic back pain **Methods of execution:** questionnaire and a structured physical examination of the back **Threshold for positive result on index test:** presence or absence of red flags from questionnaire and physical examination **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • lumbar or thoracic X-ray
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of osteoporotic vertebral fracture:** 4.5%

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			

Enthoven 2016 *(Continued)*

Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Gibson 1992

Study characteristics	
Patient Sampling	**Study design:** prospective cohort **Enrolment procedures:** consecutive sample **Sample size:** 225, 108 received imaging **Inclusion criteria:** people presenting with pain in the lumbar region of < 48 hours' duration **Exclusion criteria:** not reported
Patient characteristics and setting	**Age group:** adolescents and adults (range 16–65 years) **Gender:** not reported **Duration of pain:** < 48 hours

Gibson 1992 *(Continued)*

	History of LBP: not reported **Healthcare setting:** accident and emergency department **Country:** UK **Type of vertebral fracture:** vertebral fracture; type unclear from text **Reasons for withdrawal:** not reported
Index tests	**Index tests** • Trauma • Trauma and neurological signs • Neurological signs or straight leg raise < 40° (or both) **Methods of execution:** questionnaire **Threshold for positive result on index test:** presence or absence of red flags from questionnaire **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • X-ray; no further details
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 6.5%

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	

Gibson 1992 *(Continued)*

Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Henschke 2009

Study characteristics	
Patient Sampling	**Study design:** prospective cohort **Enrolment procedures:** consecutive sample **Sample size:** 1172 **Inclusion criteria:** people presenting with acute LBP, with < 1 week since onset **Exclusion criteria:** serious pathology diagnosed prior to consultation, and the serious pathology was considered to be the cause of the current episode of LBP
Patient characteristics and setting	**Age group:** adults (mean age 43.97 years) **Gender:** 626 men, 546 women **Duration of pain:** < 1 week: 696, 1–2 weeks: 145, 2–3 weeks: 174, 3–4 weeks: 73, 4–5 weeks: 30, 5–6 weeks 54 **History of LBP:** 888/1172 **Healthcare setting:** primary care **Country:** Australia

Henschke 2009 *(Continued)*

	Type of vertebral fracture: vertebral fracture; type unclear from text **Reasons for withdrawal:** all participants completed follow-up
Index tests	**Index tests:** 25 red flags total – 5 specific to vertebral fracture • Age > 70 years • Significant trauma • Prolonged use of corticosteroids • Altered sensation • Clinician diagnosis of fracture **Methods of execution:** checklist **Threshold for positive result on index test:** presence or absence of red flags from checklist **Experience and expertise of assessors:** unclear from text
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • clinical follow-up at 12 months with suspected cases confirmed by imaging studies and specialist review
Flow and timing	**Timing of index test and reference test:** time between index test and reference standard was 12 months
Comparative	
Notes	**Prevalence of vertebral fracture:** 0.68%

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	

Henschke 2009 *(Continued)*

Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			High
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		High risk	

Jin 2020

Study characteristics	
Patient Sampling	**Study design:** prospective observational cohort **Enrolment procedures:** consecutive sample **Sample size:** 510 **Inclusion criteria:** presented to the spine clinic with chief complaints of LBP who were postmenopausal women or women over 50 years old and men over 60 years old or people with a history of prolonged use of a glucocorticosteroid for ≥ 3 months at any age **Exclusion criteria:** back pain with pain radiating to the limbs or night/resting pain; definite recent high-energy trauma; an established diagnosis of infection, tumour, or fractures in other medical facilities before presenting to clinic; back pain with dysfunction of the central or peripheral nervous system; inability to complete the index test due to a cognitive deficit, mental impairment, or communication dysfunction (e.g. inability to hear or speak); or failure to complete an MRI scan due to any conditions, such as claustrophobia or artificial pacemaker implantation
Patient characteristics and setting	**Age group:** women > 50 years or men > 60 years (mean age 70.6 years) **Gender:** 113 men, 397 women **Duration of pain:** < 2 weeks: 219, 2–8 weeks: 138, ≥ 8 weeks: 153 **History of LBP:** not reported

Jin 2020 *(Continued)*

	Healthcare setting: hospital spine clinic **Country:** China **Type of vertebral fracture:** osteoporotic vertebral fracture **Reasons for withdrawal** • 8 participants refused to attempt the first movement because of cardiac insufficiency (2) and severe back pain (6) • MRI not performed in 56 participants for individual reasons • 8 participants with inconclusive or inconsistent MRI results were excluded from the analysis because of a suspicious malignancy/infection based on MRI
Index tests	**Index tests** • Physical examination test (Back Pain-Inducing Test) • Motion of lying to supine, rolling over, sitting up **Methods of execution:** physical examination **Threshold for positive result on index test** Defined as positive: • participants reporting back pain occurring during any movement or • participants not being able to perform any of the 3 movements due to back pain **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • MRI scan with short-tau inversion recovery (STIR) sequencing of the whole spine using a 1.5-Tesla system. Diagnoses were made by combining T1-weighted images (T1WIs), T2-weighted images (T2WIs), and STIR images in the sagittal plane.
Flow and timing	**Timing of index test and reference test:** index test and reference standard both completed within 3 days.
Comparative	
Notes	**Prevalence of vertebral fracture:** 68%

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
Could the selection of patients have introduced bias?		High risk	

Jin 2020 *(Continued)*

Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	No		
Could the patient flow have introduced bias?		High risk	

Kilic 2021

Study characteristics	
Patient Sampling	**Study design:** retrospective cohort

Kilic 2021 *(Continued)*

	Enrolment procedures: unclear from text **Sample size:** 136 **Inclusion criteria:** records of the people admitted to physical therapy and rehabilitation polyclinic with complaints of back pain **Exclusion criteria:** those whose data could not be accessed; and history of rheumatological disease, malignancy, and trauma
Patient characteristics and setting	**Age group:** adults (mean age 73.9 years) **Gender:** 17 men, 119 women **Duration of pain:** not reported **History of LBP:** not reported **Healthcare setting:** hospital (physiotherapy and rehabilitation polyclinic) **Country:** Turkey **Type of vertebral fracture:** spontaneous osteoporosis vertebral fracture **Reasons for withdrawal:** all participants included in analysis
Index tests	**Index tests** • Aged ≥ 75 years • Aged 65–75 years **Methods of execution:** hospital records system **Threshold for positive result on index test:** aged ≥ 75 years and aged 65–75 years **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • radiograph; presence of fractures in participants; T4–L5 in lateral spinal radiographs anterior, middle, or posterior heights in intervertebrae by semi-quantitative method or lateral height in ≥ 1 vertebra between T4 and L5
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 54.4%

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Unclear		

Kilic 2021 *(Continued)*

Did the study avoid inappropriate exclusions?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Patrick 1983

Study characteristics	
Patient Sampling	**Study design:** retrospective chart review **Enrolment procedures:** consecutive sample, reason for patient presentation not clear **Sample size:** 552

Patrick 1983 *(Continued)*

	Inclusion criteria: lumbar spine series ordered in the emergency departments were enroled **Exclusion criteria:** unclear from text
Patient characteristics and setting	**Age group:** children to adults (range 6–95 years) **Gender:** 296 male, 256 female **Duration of pain:** unclear from text **History of LBP:** unclear from text **Healthcare setting:** accident and emergency department and university hospital medical centre **Country:** USA **Type of vertebral fracture:** vertebral compression fracture **Reasons for withdrawal:** all participants included in analysis
Index tests	**Index tests** • Trauma • Tenderness • LBP with radiation or hip pain, or both • Spasm • Sensory deficit • Motor deficit • Tendon reflex abnormality • Positive straight leg raise • Contusion/abrasion • Multiple findings **Methods of execution:** hospital records system **Threshold for positive result on index test:** presence or absence of red flags from checklist **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • X-ray; not further defined
Flow and timing	**Timing of index test and reference standard:** index test and reference standard performed at time of admission
Comparative	
Notes	**Prevalence of vertebral fracture:** 7.2%

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			

Patrick 1983 *(Continued)*

Was a consecutive or random sample of patients enrolled?	Unclear		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Unclear		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Premkumar 2018

Study characteristics	
Patient Sampling	**Study design:** retrospective cohort **Enrolment procedures:** consecutive sample **Sample size:** 9940

Premkumar 2018 *(Continued)*

	Inclusion criteria: people who presented as a new patient with a chief complaint of LBP with or without leg pain **Exclusion criteria:** chief complaint of deformity rather than pain, chief complaint of neck pain
Patient characteristics and setting	**Age group:** adults (mean age 56.8 years) **Gender:** 5102 men, 4838 women **Duration of pain:** not reported **History of LBP:** not reported **Healthcare setting:** academic multidisciplinary spine centre **Country:** USA **Type of vertebral fracture:** vertebral fracture; type unclear from text **Reasons for withdrawal:** all participants included in analysis
Index tests	**Index tests** • Age > 50 years • Age > 70 years • Trauma • Combination 1: trauma and age > 50 years • Combination 2: trauma and age > 70 years **Methods of execution:** questionnaire **Threshold for positive result on index test:** presence or absence of red flags from questionnaire **Experience and expertise of assessors:** fellowship trained spine surgeons
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • imaging; type of imaging not specified. Diagnostic information was drawn directly from the physician entry and was corroborated by imaging reports.
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 5.6%

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		

Premkumar 2018 *(Continued)*

Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Unclear		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Reinus 1998

Study characteristics	
Patient Sampling	**Study design:** prospective cohort **Enrolment procedures:** unclear from text **Sample size:** 482 **Inclusion criteria:** all patients receiving lumbosacral radiographs in a 14-month period

Reinus 1998 *(Continued)*

	Exclusion criteria: unclear from text
Patient characteristics and setting	**Age group:** adolescents and adults (range 17–98 years) **Gender:** 168 male, 314 female **Duration of pain:** unclear from text **History of LBP:** unclear from text **Healthcare setting:** accident and emergency department **Country:** USA **Type of vertebral fracture:** vertebral compression fracture **Reasons for withdrawal:** unclear from text
Index tests	**Index test(s)** • Trauma • Neurological deficit • Trauma or neoplasm (indeterminate age fracture) **Methods of execution:** questionnaire **Threshold for positive result on index test:** presence or absence of red flags from questionnaire **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • X-ray – lumbosacral anteroposterior, lateral, bilateral posterior oblique, and coned down radiological views lumbar views
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of fracture:** 11%

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Unclear		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			

Reinus 1998 *(Continued)*

Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Roman 2010

Study characteristics	
Patient Sampling	**Study design:** retrospective chart review **Enrolment procedures:** consecutive sample **Sample size:** 1448 **Inclusion criteria:** all patients with a lumbar-related disorder **Exclusion criteria:** unclear from text
Patient characteristics and setting	**Age group:** adults **Gender:** 587 men, 861 women **Duration of pain:** not reported **History of LBP:** not reported

Roman 2010 *(Continued)*

	Healthcare setting: university; Department of Surgery **Country:** USA **Type of vertebral fracture:** osteoporotic vertebral compression fracture **Reasons for withdrawal:** all participants included in analysis
Index tests	**Index tests** • Gender • Age • BMI • Gait abnormality • Regular exercise • Sitting pain • Osteoarthritis • Leg or buttock pain • Multiple signs **Methods of execution:** clinical examination **Threshold for positive result on index test:** presence or absence of red flags from during physical examination **Experience and expertise of assessors:** 2 assessors (4 and 22 years' experience)
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • standard radiograph or CT; assessing sagittal alignment, vertebral body compression, and spinal canal dimensions
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 3%

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			

Roman 2010 *(Continued)*

Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		High risk	

Scavone 1981

Study characteristics	
Patient Sampling	**Study design:** retrospective chart review **Enrolment procedures:** unclear from text **Sample size:** 871 **Inclusion criteria:** people with lumbar spine X-ray **Exclusion criteria:** unclear from text
Patient characteristics and setting	**Age group:** unclear from text **Gender:** unclear from text **Duration of pain:** unclear from text **History of LBP:** unclear from text

Scavone 1981 *(Continued)*

	Healthcare setting: university teaching hospital medical centre **Country:** USA **Type of vertebral fracture:** vertebral fracture; unclear from text **Reasons for withdrawal:** unclear from text; history for 57 patients could not be obtained
Index tests	**Index tests** • Major trauma • Minor trauma • Tenderness • LBP with radiation • Hip/leg pain • Muscle spasm • Neurological deficits • Sciatica • Abnormal physical examination **Methods of execution:** chart **Threshold for positive result on index test:** presence or absence of red flags from chart **Experience and expertise of assessors:** unclear from text
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on • AP and lateral X-ray views
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 3%

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Unclear		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		

Scavone 1981 *(Continued)*

If a threshold was used, was it pre-specified?	Unclear		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Unclear		
Could the patient flow have introduced bias?		Unclear risk	

Sugiura 2021

Study characteristics	
Patient Sampling	**Study design:** retrospective comparative cohort study **Enrolment procedures:** consecutive sample **Sample size:** 101 **Inclusion criteria:** adolescent presenting for rehabilitation within 1 month of acute LBP **Exclusion criteria:** participants with lower extremity symptoms (to exclude the possibility of radicular back pain), clear spondylolysis or spondylolisthesis based on plain radiography findings, and other spinal disorders based on MRI findings
Patient characteristics and setting	**Age group:** adolescents aged < 18 years (mean age 14.4 years; range 10–18 years) **Gender:** 74 male, 27 female **Duration of pain:** < 1 month

Sugiura 2021 *(Continued)*

	History of LBP: not reported **Healthcare setting:** unclear from text **Country:** Japan **Type of vertebral fracture:** early-stage spondylolysis; vertebral stress fracture **Reasons for withdrawal:** unclear from text; participants missing from analysis
Index tests	**Index tests** • Gender • Participation in sports activities • Presence or absence of LBP during lumbar spine extension in a standing position (the hyperextension test) • Presence or absence of LBP during lumbar spine flexion in a standing position (the hyperflexion test) • Pain quality (sharp or dull) • Pain extent (categorised as a fingertip-sized or palm-sized area) • Pain location (pain on the left or right side (or both) or central pain centre) **Methods of execution:** physical examination **Threshold for positive result on index test:** presence or absence of red flags during physical examination **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference tests:** vertebral fracture on: • MRI; the diagnosis was made when the lumbar spine pedicle showed high signal intensity on fat saturation T2-weighted images and low signal intensity on T1-weighted images
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 52.5%

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern

Sugiura 2021 *(Continued)*

DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Therriault 2020

Study characteristics	
Patient Sampling	**Study design:** retrospective case series **Enrolment procedures:** consecutive sample **Sample size:** 1025 **Inclusion criteria:** people with a primary complaint of LBP **Exclusion criteria:** person did not participate in an organised sport, defined as physical activity with formal practice or competition ≥ 2 times per week, or

Therriault 2020 *(Continued)*

	diagnosis of spondylolysis was not substantiated by imaging. Any patient without a report of LBP was removed from the potential patient list.
Patient characteristics and setting	**Age group:** adolescents (mean age 15 years) **Gender:** 455 male, 570 female **Duration of pain:** mean duration 19.5 weeks **History of LBP:** 253/1025 participants **Healthcare setting:** sports medicine department at a large children's hospital **Country:** USA **Type of vertebral fracture:** active spondylolysis (vertebral stress fracture) **Reasons for withdrawal:** all participants included in analysis
Index tests	**Index tests** • Physical examination and history • Sex • Previous history of LBP • Pain at rest • Pain reported with lumbar extension • Pain reported with lumbar flexion • Pain reported with lumbar rotation • Pain reported with lumbar lateral flexion • Radicular symptoms reported • Numbness or tingling reported • Single-leg Hyperextension Test • Vertebral tenderness **Methods of execution:** physical examination and history **Threshold for positive result on index test:** presence or absence of red flags during physical examination and history **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • anteroposterior and standing lateral radiographs; if radiographs did not identify a spondylolytic injury and the physician suspected this diagnosis further imaging was performed; • MRI; or • SPECT.
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 22%

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns

Therriault 2020 *(Continued)*

DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		High risk	

van den Bosch 2004

Study characteristics	
Patient Sampling	**Study design:** retrospective chart review **Enrolment procedures:** random sample **Sample size:** 2007 **Inclusion criteria:** people with full radiographic and demographic details **Exclusion criteria:** unclear from text
Patient characteristics and setting	**Age group:** adolescents and adults (range 0 to > 75 years) (age range as reported by study authors) **Gender:** 845 male, 1162 female **Duration of pain:** not reported **History of LBP:** not reported **Healthcare setting:** hospital setting **Country:** UK **Type of vertebral fracture:** osteoporotic and traumatic fracture; data not differentiated for different type of fracture **Reasons for withdrawal:** not reported
Index tests	**Index tests** • Age • Gender **Methods of execution:** hospital records system **Threshold for positive result on index test:** presence or absence of red flags from hospital record system **Experience and expertise of assessors:** not reported
Target condition and reference standard(s)	**Reference test(s):** vertebral fracture on: • X-ray; not further defined
Flow and timing	**Timing of index test and reference test:** unclear from text
Comparative	
Notes	**Prevalence of vertebral fracture:** 4.1%

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		

van den Bosch 2004 *(Continued)*

Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (Index Test)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

AP: anterior-posterior; BMI: body mass index; CT: computed tomography; GP: general practitioner; LBP: low back pain; MRI: magnetic resonance imaging; SPECT: single-photon emission computerised tomography.

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Althoff 2017	Inappropriate population sample investigated
Baleanu 2020	Inappropriate population sample investigated

Study	Reason for exclusion
Bian 2022	People already diagnosed with vertebral fracture
Borgen 2020	Inappropriate population sample investigated
Bottai 2016	People already diagnosed with vertebral fracture
Cheung 2018	Inappropriate study design
Clark 2017	Inappropriate study design
Della-Giustina 2013	Inappropriate study design
Della-Giustina 2015	Inappropriate study design
Durham 1995	Inappropriate population sample investigated
Enoki 2022	Inappropriate population sample investigated
Finucane 2020	Inappropriate study design
Frankel 1994	Inappropriate population sample investigated
Gaddikeri 2018	No index test assessed
Galliker 2019	Wrong study design
Gazzola 2015	Inappropriate population sample investigated
Gestring 2002	Inappropriate population sample investigated
Gill 2013	Inappropriate population sample investigated
Hiett 2014	No appropriate data available
Holmes 2003	Inappropriate population sample investigated
Hsu 2003	Inappropriate population sample investigated
Johannesdottir 2021	Inappropriate study design
Johansson 2018	Inappropriate population sample investigated
Kendler 2016	Inappropriate study design
Khera 2022	No appropriate data available
Kherad 2015	Inappropriate population sample investigated
Korner 2014	Inappropriate study design
Kuru 2014	Inappropriate population sample investigated
Manji 2022	People already diagnosed with vertebral fracture
Mazziotti 2016	Inappropriate population sample investigated

Study	Reason for exclusion
Min 2017	Inappropriate study design
Mizukami 2014	No index test assessed
Moretti 2019	Inappropriate population sample investigated
Padlina 2017	Inappropriate population sample investigated
Pressney 2014	Inappropriate study design
Rajakulasingam 2022	Inappropriate population sample investigated
Rapala 2012	No appropriate data available
Reito 2018	No appropriate data available
Rudman 2019	Inappropriate population sample investigated
Sakai 2016	Inappropriate population sample investigated
Samuels 1993	Inappropriate population sample investigated
Scaturro 2020	No appropriate data available
Shah 2021	Inappropriate population sample investigated
Soto-Subiabre 2020	People already diagnosed with vertebral fracture
St Jeor 2020	Inappropriate population sample investigated
Sucato 2012	Inappropriate study design
Terakado 2017	No index test assessed
Terregino 1995	Inappropriate population sample investigated
Trout 2015	Inappropriate study design
Upadhye 2016	Inappropriate study design
van der Jagt-Willems 2012	Inappropriate population sample investigated
Yoshimura 2017	Inappropriate population sample investigated
Zong 2016	Inappropriate population sample investigated
Zorin 2020	Inappropriate population sample investigated

LBP: low back pain.

DATA

Presented below are all the data for all of the tests entered into the review.

Table Tests. Data tables by test

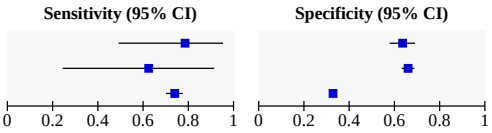
Test	No. of studies	No. of participants
1 Age > 50 years	3	11423
2 Age > 54 years	2	3179
3 Age > 52 years	1	1449
4 Age > 64 years	2	3179
5 Age > 70 years	2	11112
6 Age > 74 years	2	3179
7 Female gender	5	5250
8 Female > 54 years	2	3179
9 Female > 64 years	2	3182
10 Female > 74 years	2	3280
11 Trauma	8	14105
12 Corticosteroid use	3	2152
13 OA	2	2117
14 No regular exercise	1	1448
15 Sensation change	4	3620
16 Motor deficit	2	1424
17 DTR abnormality	2	1423
18 Neurological signs	2	590
19 SLR	1	552
20 Tenderness	3	2427
21 Spasm	2	1423
22 Contusion/abrasion	1	552
23 Sciatica	2	1896
24 Hip/leg pain	1	871
25 Absence of buttock/leg pain	1	1448
26 BMI < 23	1	1448

Test	No. of studies	No. of participants
27 Decreased pain on sitting	1	1448
28 No gait abnormality	1	1448
29 Trauma and neurological signs	1	108
30 Multiple findings	1	552
31 Henschke rule 1 sign positive	1	1172
32 Henschke rule 2 positive signs	1	1172
33 Henschke rule 3 positive signs	1	1172
34 1 of 5 positive - Roman Cluster	1	1449
35 2 of 5 positive - Roman Cluster	1	1449
36 3 of 5 positive - Roman Cluster	1	1448
37 4 of 5 positive - Roman Cluster	1	1448
38 5 of 5 positive - Roman Cluster	1	1448
39 Age ≥ 75 years	2	805
40 Osteoporosis	1	669
41 Percussion tenderness of spine	1	669
42 Sudden decrease in height	1	669
43 Severe disability	1	669
44 Acute onset of pain	1	669
45 Back pain score ≥ 7	1	669
46 Thoracic back pain	1	669
47 BPIT	1	510
48 Age 65–75 years	1	136
49 Participation in sports	1	100
50 Hyperextension	2	1111
51 Hyperflexion	2	1109
52 Sharp pain	1	82
53 Pain extent (fingertip size)	1	96
54 Pain location	1	99

Test	No. of studies	No. of participants
55 Previous history of LBP	1	1025
56 Waking night pain	1	1025
57 Lumbar rotation	1	1010
58 Lumbar lateral flexion	1	1012
59 Single leg hyperextension	1	989
60 2 of 3 negative - Therriault cluster	0	0
61 2 of 3 positive - Therriault cluster	0	0
62 3 of 3 negative - Therriault cluster	0	0

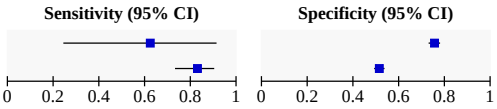
Test 1. Age > 50 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Deyo 1986	11	108	3	189	0.79 [0.49, 0.95]	0.64 [0.58, 0.69]
Henschke 2009	5	395	3	769	0.63 [0.24, 0.91]	0.66 [0.63, 0.69]
Premkumar 2018	410	6300	144	3086	0.74 [0.70, 0.78]	0.33 [0.32, 0.34]



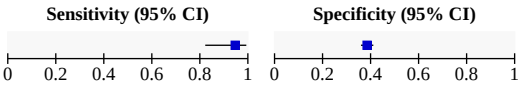
Test 2. Age > 54 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	5	283	3	881	0.63 [0.24, 0.91]	0.76 [0.73, 0.78]
van den Bosch 2004	69	932	14	992	0.83 [0.73, 0.90]	0.52 [0.49, 0.54]



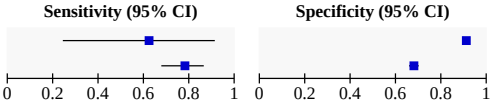
Test 3. Age > 52 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	36	867	2	544	0.95 [0.82, 0.99]	0.39 [0.36, 0.41]



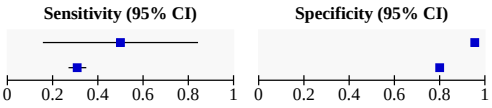
Test 4. Age > 64 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	5	102	3	1062	0.63 [0.24, 0.91]	0.91 [0.89, 0.93]
van den Bosch 2004	65	613	18	1311	0.78 [0.68, 0.87]	0.68 [0.66, 0.70]



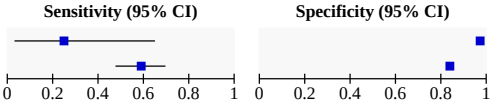
Test 5. Age > 70 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	4	52	4	1112	0.50 [0.16, 0.84]	0.96 [0.94, 0.97]
Premkumar 2018	171	1875	383	7511	0.31 [0.27, 0.35]	0.80 [0.79, 0.81]



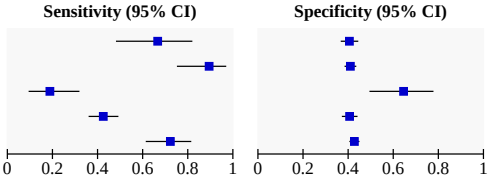
Test 6. Age > 74 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	2	31	6	1133	0.25 [0.03, 0.65]	0.97 [0.96, 0.98]
van den Bosch 2004	49	308	34	1616	0.59 [0.48, 0.70]	0.84 [0.82, 0.86]



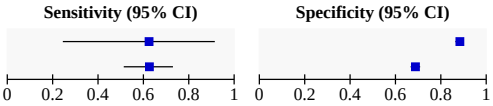
Test 7. Female gender

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	22	378	11	258	0.67 [0.48, 0.82]	0.41 [0.37, 0.44]
Roman 2010	34	832	4	578	0.89 [0.75, 0.97]	0.41 [0.38, 0.44]
Sugiura 2021	10	17	43	31	0.19 [0.09, 0.32]	0.65 [0.49, 0.78]
Therriault 2020	97	473	131	324	0.43 [0.36, 0.49]	0.41 [0.37, 0.44]
van den Bosch 2004	60	1101	23	823	0.72 [0.61, 0.82]	0.43 [0.41, 0.45]



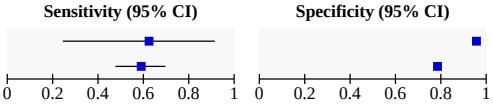
Test 8. Female > 54 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	5	135	3	1029	0.63 [0.24, 0.91]	0.88 [0.86, 0.90]
van den Bosch 2004	52	600	31	1324	0.63 [0.51, 0.73]	0.69 [0.67, 0.71]



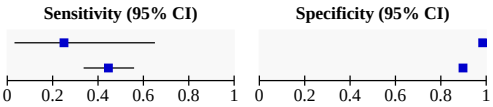
Test 9. Female > 64 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	5	50	3	1117	0.63 [0.24, 0.91]	0.96 [0.94, 0.97]
van den Bosch 2004	49	413	34	1511	0.59 [0.48, 0.70]	0.79 [0.77, 0.80]



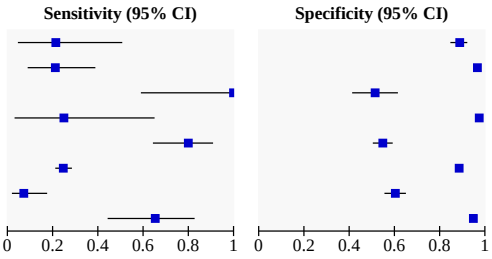
Test 10. Female > 74 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	2	18	6	1146	0.25 [0.03, 0.65]	0.98 [0.98, 0.99]
van den Bosch 2004	37	207	46	1818	0.45 [0.34, 0.56]	0.90 [0.88, 0.91]



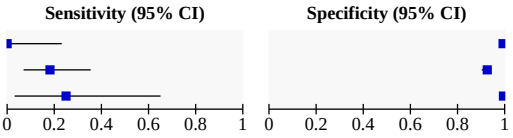
Test 11. Trauma

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Deyo 1986	3	33	11	264	0.21 [0.05, 0.51]	0.89 [0.85, 0.92]
Enthoven 2016	7	21	26	615	0.21 [0.09, 0.39]	0.97 [0.95, 0.98]
Gibson 1992	7	49	0	52	1.00 [0.59, 1.00]	0.51 [0.41, 0.62]
Henschke 2009	2	29	6	1135	0.25 [0.03, 0.65]	0.98 [0.96, 0.98]
Patrick 1983	32	231	8	281	0.80 [0.64, 0.91]	0.55 [0.50, 0.59]
Premkumar 2018	137	1067	417	8319	0.25 [0.21, 0.29]	0.89 [0.88, 0.89]
Reinus 1998	4	169	51	258	0.07 [0.02, 0.18]	0.60 [0.56, 0.65]
Scavone 1981	17	43	9	802	0.65 [0.44, 0.83]	0.95 [0.93, 0.96]



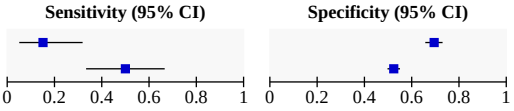
Test 12. Corticosteroid use

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Deyo 1986	0	2	14	295	0.00 [0.00, 0.23]	0.99 [0.98, 1.00]
Enthoven 2016	6	47	27	589	0.18 [0.07, 0.35]	0.93 [0.90, 0.95]
Henschke 2009	2	6	6	1158	0.25 [0.03, 0.65]	0.99 [0.99, 1.00]



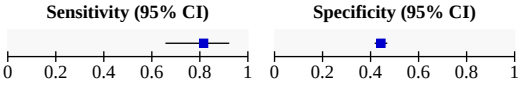
Test 13. OA

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	5	194	28	442	0.15 [0.05, 0.32]	0.69 [0.66, 0.73]
Roman 2010	19	671	19	739	0.50 [0.33, 0.67]	0.52 [0.50, 0.55]



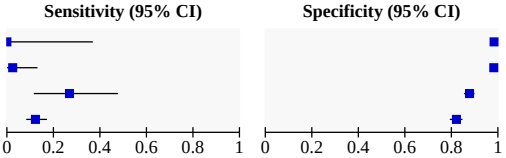
Test 14. No regular exercise

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	31	785	7	625	0.82 [0.66, 0.92]	0.44 [0.42, 0.47]



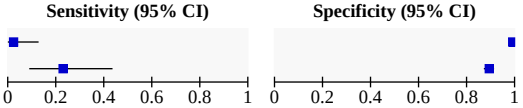
Test 15. Sensation change

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	0	19	8	1145	0.00 [0.00, 0.37]	0.98 [0.97, 0.99]
Patrick 1983	1	9	39	503	0.03 [0.00, 0.13]	0.98 [0.97, 0.99]
Scavone 1981	7	103	19	742	0.27 [0.12, 0.48]	0.88 [0.85, 0.90]
Therriault 2020	28	142	200	655	0.12 [0.08, 0.17]	0.82 [0.79, 0.85]



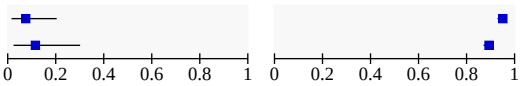
Test 16. Motor deficit

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Patrick 1983	1	4	40	508	0.02 [0.00, 0.13]	0.99 [0.98, 1.00]
Scavone 1981	6	89	20	756	0.23 [0.09, 0.44]	0.89 [0.87, 0.91]



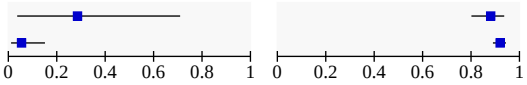
Test 17. DTR abnormality

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Patrick 1983	3	25	37	487	0.07 [0.02, 0.20]	0.95 [0.93, 0.97]		
Scavone 1981	3	90	23	755	0.12 [0.02, 0.30]	0.89 [0.87, 0.91]		



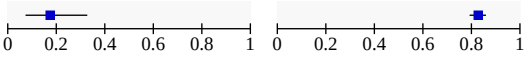
Test 18. Neurological signs

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Gibson 1992	2	12	5	89	0.29 [0.04, 0.71]	0.88 [0.80, 0.94]		
Reinus 1998	3	34	52	393	0.05 [0.01, 0.15]	0.92 [0.89, 0.94]		



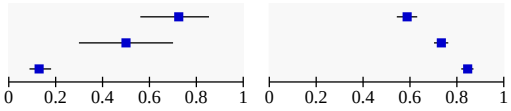
Test 19. SLR

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Patrick 1983	7	88	33	424	0.17 [0.07, 0.33]	0.83 [0.79, 0.86]		



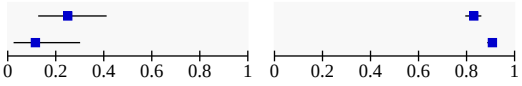
Test 20. Tenderness

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Patrick 1983	29	211	11	301	0.72 [0.56, 0.85]	0.59 [0.54, 0.63]		
Scavone 1981	13	225	13	620	0.50 [0.30, 0.70]	0.73 [0.70, 0.76]		
Therriault 2020	29	120	195	660	0.13 [0.09, 0.18]	0.85 [0.82, 0.87]		



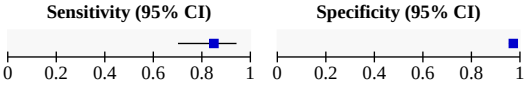
Test 21. Spasm

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Patrick 1983	10	87	30	425	0.25 [0.13, 0.41]	0.83 [0.79, 0.86]		
Scavone 1981	3	78	23	767	0.12 [0.02, 0.30]	0.91 [0.89, 0.93]		



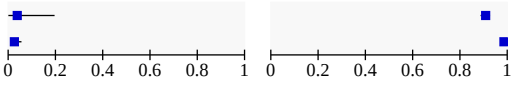
Test 22. Contusion/abrasion

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Patrick 1983	34	14	6	498	0.85 [0.70, 0.94]	0.97 [0.95, 0.98]		



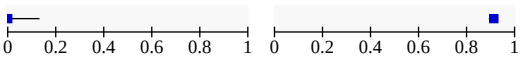
Test 23. Sciatica

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Scavone 1981	1	77	25	768	0.04 [0.00, 0.20]	0.91 [0.89, 0.93]		
Therriault 2020	6	11	222	786	0.03 [0.01, 0.06]	0.99 [0.98, 0.99]		



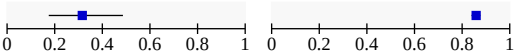
Test 24. Hip/leg pain

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Scavone 1981	0	73	26	772	0.00 [0.00, 0.13]	0.91 [0.89, 0.93]		



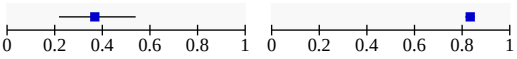
Test 25. Absence of buttock/leg pain

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	12	199	26	1211	0.32 [0.18, 0.49]	0.86 [0.84, 0.88]		



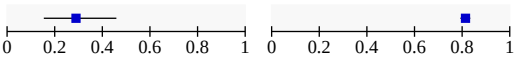
Test 26. BMI < 23

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	14	234	24	1176	0.37 [0.22, 0.54]	0.83 [0.81, 0.85]		



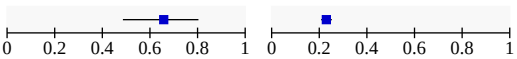
Test 27. Decreased pain on sitting

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	11	262	27	1148	0.29 [0.15, 0.46]	0.81 [0.79, 0.83]		



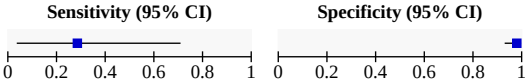
Test 28. No gait abnormality

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	25	1086	13	324	0.66 [0.49, 0.80]	0.23 [0.21, 0.25]		



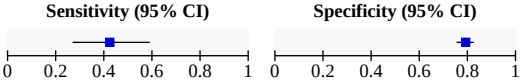
Test 29. Trauma and neurological signs

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Gibson 1992	2	2	5	99	0.29 [0.04, 0.71]	0.98 [0.93, 1.00]		



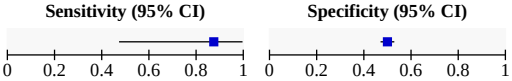
Test 30. Multiple findings

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Patrick 1983	17	106	23	406	0.42 [0.27, 0.59]	0.79 [0.76, 0.83]		



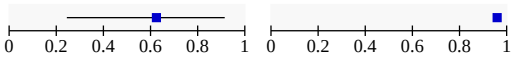
Test 31. Henschke rule 1 sign positive

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	7	582	1	582	0.88 [0.47, 1.00]	0.50 [0.47, 0.53]		



Test 32. Henschke rule 2 positive signs

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	5	47	3	1117	0.63 [0.24, 0.91]	0.96 [0.95, 0.97]		

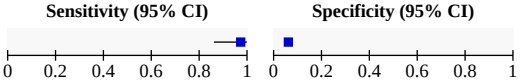


Test 33. Henschke rule 3 positive signs

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Henschke 2009	3	0	5	1164	0.38 [0.09, 0.76]	1.00 [1.00, 1.00]		

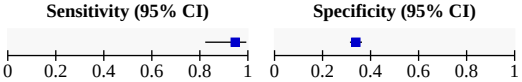
Test 34. 1 of 5 positive - Roman Cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	37	1324	1	87	0.97 [0.86, 1.00]	0.06 [0.05, 0.08]		



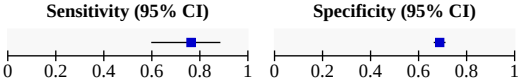
Test 35. 2 of 5 positive - Roman Cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	36	933	2	478	0.95 [0.82, 0.99]	0.34 [0.31, 0.36]		



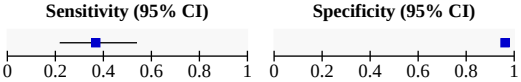
Test 36. 3 of 5 positive - Roman Cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	29	440	9	970	0.76 [0.60, 0.89]	0.69 [0.66, 0.71]		



Test 37. 4 of 5 positive - Roman Cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	14	54	24	1356	0.37 [0.22, 0.54]	0.96 [0.95, 0.97]		

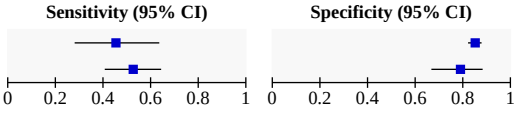


Test 38. 5 of 5 positive - Roman Cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Roman 2010	1	5	36	1406	0.03 [0.00, 0.14]	1.00 [0.99, 1.00]		

Test 39. Age ≥ 75 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	15	94	18	542	0.45 [0.28, 0.64]	0.85 [0.82, 0.88]		
Kilic 2021	39	13	35	49	0.53 [0.41, 0.64]	0.79 [0.67, 0.88]		

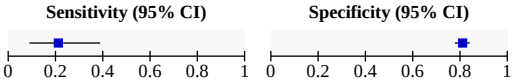


Test 40. Osteoporosis

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	12	74	21	562	0.36 [0.20, 0.55]	0.88 [0.86, 0.91]

Test 41. Percussion tenderness of spine

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	7	120	26	516	0.21 [0.09, 0.39]	0.81 [0.78, 0.84]

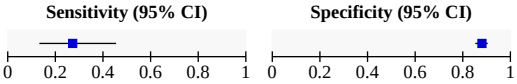


Test 42. Sudden decrease in height

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	3	20	30	616	0.09 [0.02, 0.24]	0.97 [0.95, 0.98]

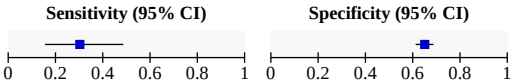
Test 43. Severe disability

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	9	76	24	560	0.27 [0.13, 0.46]	0.88 [0.85, 0.90]



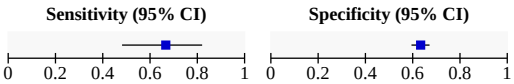
Test 44. Acute onset of pain

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	10	222	23	414	0.30 [0.16, 0.49]	0.65 [0.61, 0.69]



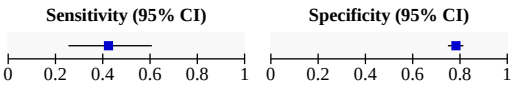
Test 45. Back pain score ≥ 7

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	22	233	11	403	0.67 [0.48, 0.82]	0.63 [0.59, 0.67]



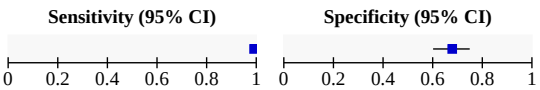
Test 46. Thoracic back pain

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Enthoven 2016	14	138	19	498	0.42 [0.25, 0.61]	0.78 [0.75, 0.81]



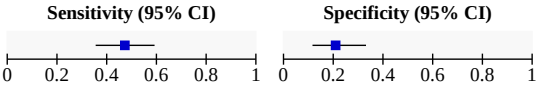
Test 47. BPIT

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Jin 2020	342	53	3	112	0.99 [0.97, 1.00]	0.68 [0.60, 0.75]



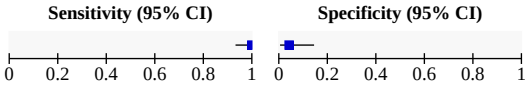
Test 48. Age 65–75 years

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Kilic 2021	35	49	39	13	0.47 [0.36, 0.59]	0.21 [0.12, 0.33]



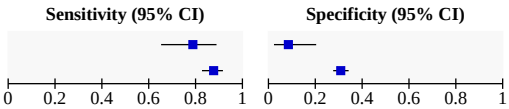
Test 49. Participation in sports

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Sugiura 2021	53	45	0	2	1.00 [0.93, 1.00]	0.04 [0.01, 0.15]



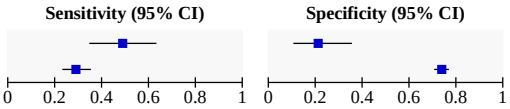
Test 50. Hyperextension

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Sugiura 2021	41	43	11	4	0.79 [0.65, 0.89]	0.09 [0.02, 0.20]
Therriault 2020	200	542	28	242	0.88 [0.83, 0.92]	0.31 [0.28, 0.34]



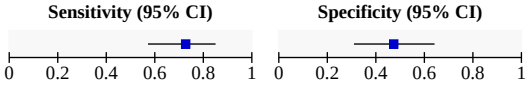
Test 51. Hyperflexion

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Sugiura 2021	25	37	26	10	0.49 [0.35, 0.63]	0.21 [0.11, 0.36]
Therriault 2020	66	204	161	580	0.29 [0.23, 0.35]	0.74 [0.71, 0.77]



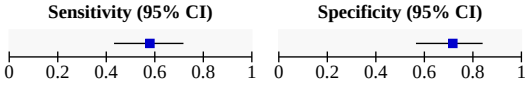
Test 52. Sharp pain

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Sugiura 2021	32	20	12	18	0.73 [0.57, 0.85]	0.47 [0.31, 0.64]



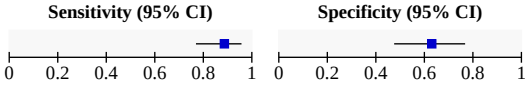
Test 53. Pain extent (fingertip size)

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Sugiura 2021	29	13	21	33	0.58 [0.43, 0.72]	0.72 [0.57, 0.84]



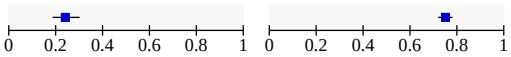
Test 54. Pain location

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Sugiura 2021	47	17	6	29	0.89 [0.77, 0.96]	0.63 [0.48, 0.77]



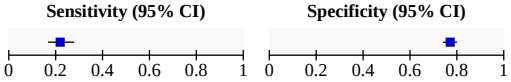
Test 55. Previous history of LBP

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Therriault 2020	55	198	173	599	0.24 [0.19, 0.30]	0.75 [0.72, 0.78]		



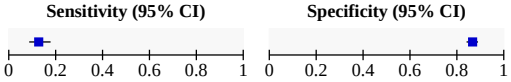
Test 56. Waking night pain

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Therriault 2020	50	182	178	615	0.22 [0.17, 0.28]	0.77 [0.74, 0.80]		



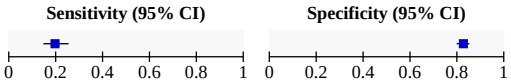
Test 57. Lumbar rotation

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Therriault 2020	29	104	198	679	0.13 [0.09, 0.18]	0.87 [0.84, 0.89]		



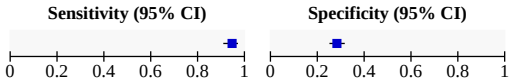
Test 58. Lumbar lateral flexion

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Therriault 2020	45	135	183	649	0.20 [0.15, 0.26]	0.83 [0.80, 0.85]		



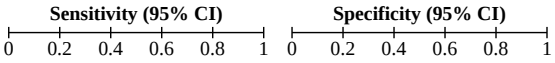
Test 59. Single leg hyperextension

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Therriault 2020	216	545	12	216	0.95 [0.91, 0.97]	0.28 [0.25, 0.32]		



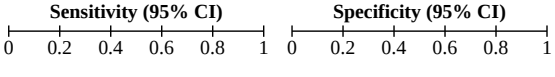
Test 60. 2 of 3 negative - Therriault cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)



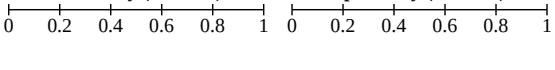
Test 61. 2 of 3 positive - Therriault cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)



Test 62. 3 of 3 negative - Therriault cluster

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)



ADDITIONAL TABLES

Table 1. Criteria for study-specific judgement of the informativeness of tests based on positive likelihood ratios

Test result (Deeks 2004; Jaeschke 1994)	**Interpretation**
Upper bound of confidence interval < 2	Uninformative test
Lower bound of confidence interval spans 1 but upper bound > 2	Indeterminate result because we have an imprecise estimate of the test accuracy
Point estimate in the range 2–5 and lower bound of confidence interval is > 1.0	Small increase in likelihood of fracture
Point estimate in the range 5–10 and lower bound of confidence interval is > 1.0	Moderate increase in likelihood of fracture
Point estimate > 10 and lower bound of confidence interval is > 1.0	Large increase in likelihood of fracture

Table 2. Results of all red flags taken during a history in primary care

Study	**Index test**	**Reference standard**	**Sample size**	**Prevalence**	**+LR (95% CI)**	**−LR (95% CI)**
Deyo 1986	Age > 50 years	X-ray	621	0.045	2.16 (1.58 to 2.95)	0.34 (0.12 to 0.92)
Henschke 2009	Age > 50 years	Follow-up/ imaging	1178	0.007	1.84 (1.07 to 3.17)	0.57 (0.23 to 1.39)
van den Bosch 2004	Age > 54 years	X-ray	2100	0.041	1.72 (1.54 to 1.91)	0.33 (0.20 to 0.53)
Henschke 2009	Age > 54 years	Follow-up/ imaging	1178	0.007	2.57 (1.49 to 4.44)	0.50 (0.2 to 1.21)
van den Bosch 2004	Age > 64 years	X-ray	2100	0.041	2.46 (2.16 to 2.8)	0.32 (0.21 to 0.48)
Henschke 2009	Age > 64 years	Follow-up/ imaging	1178	0.007	7.13 (4.04 to 12.59)	0.41 (0.17 to 1.01)
Henschke 2009	Age > 70 years	Follow-up/ imaging	1178	0.007	11.19 (5.33 to 23.51)	0.52 (0.26 to 1.05)
van den Bosch 2004	Age > 74 years	X-ray	2100	0.041	3.69 (3.00 to 4.53)	0.49 (0.38 to 0.63)
Henschke 2009	Age > 74 years	Follow-up/ imaging	1178	0.007	9.39 (2.69 to 32.75)	0.77 (0.52 to 1.15)
Enthoven 2016	Age ≥ 75 years	X-ray	669	0.045	3.08 (2.03 to 4.67)	0.64 (0.47 to 0.88)
van den Bosch 2004	Female gender	X-ray	2100	0.041	1.26 (1.10 to 1.45)	0.65 (0.46 to 0.92)
Enthoven 2016	Female gender	X-ray	669	0.045	1.12 (0.87 to 1.44)	0.82 (0.50 to 1.34)
Therriault 2020	Female gender	X-ray, MRI, or SPECT	1025	0.22	0.72 (0.61 to 0.84)	1.41 (1.23 to 1.63)

Table 2. Results of all red flags taken during a history in primary care *(Continued)*

van den Bosch 2004	Female > 54 years	X-ray	2100	0.041	2.01 (1.68 to 2.40)	0.54 (0.41 to 0.72)
Henschke 2009	Female > 54 years	Follow-up/ imaging	1178	0.007	5.39 (3.08 to 9.43)	0.42 (0.17 to 1.04)
van den Bosch 2004	Female > 64 years	X-ray	2100	0.041	2.75 (2.26 to 3.35)	0.52 (0.40 to 0.68)
Henschke 2009	Female > 64 years	Follow-up/ imaging	1178	0.007	14.59 (8.00 to 26.61)	0.39 (0.16 to 0.96)
van den Bosch 2004	Female > 74 years	X-ray	2100	0.041	4.14 (3.17 to 5.44)	0.62 (0.51 to 0.75)
Henschke 2009	Female > 74 years	Follow-up/ imaging	1178	0.007	16.17 (4.47 to 58.43)	0.76 (0.51 to 1.14)
Deyo 1986	Significant trauma	X-ray	621	0.045	1.93 (0.67 to 5.53)	0.88 (0.67 to 1.17)
Henschke 2009	Significant trauma	Follow-up/ imaging	1178	0.007	10.03 (2.87 to 35.13)	0.77 (0.52 to 1.15)
Scavone 1981	Significant trauma	X-ray	871	0.030	12.85 (8.58 to 19.24)	0.36 (0.22 to 0.62)
Enthoven 2016	Trauma	X-ray	669	0.045	6.42 (2.94 to 14.02)	0.82 (0.68 to 0.97)
Deyo 1986	Corticosteroid use	X-ray	621	0.045	3.97 (0.20 to 79.15)	0.98 (0.89 to 1.07)
Enthoven 2016	Prolonged corti-costeroid use	X-ray	669	0.045	2.46 (1.13 to 5.34)	0.88 (0.75 to 1.04)
Henschke 2009	Prolonged corti-costeroid use	Follow-up/ imaging	1178	0.007	48.50 (11.46 to 204.98)	0.75 (0.51 to 1.13)
Enthoven 2016	Osteoporosis	X-ray	669	0.045	3.13 (1.90 to 5.15)	0.72 (0.56 to 0.93)
Enthoven 2016	Acute onset of pain	X-ray	669	0.045	0.87 (0.52 to 1.47)	1.07 (0.85 to 1.35)
Enthoven 2016	Back pain intensi-ty score ≥ 7	X-ray	669	0.045	1.82 (1.40 to 2.37)	0.53 (0.32 to 0.86)
Therriault 2020	Previous history of LBP	X-ray, MRI, or SPECT	1025	0.22	0.97 (0.75 to 1.26)	1.01 (0.93 to 1.10)
Therriault 2020	Waking night pain	X-ray, MRI, or SPECT	1025	0.22	0.96 (0.73 to 1.27)	1.01 (0.94 to 1.09)
Enthoven 2016	Osteoarthritis in hip/knee	X-ray	669	0.045	0.50 (0.22 to 1.12)	1.22 (1.05 to 1.42)
Enthoven 2016	Thoracic back pain	X-ray	669	0.045	1.96 (1.28 to 2.99)	0.74 (0.55 to 0.99)

+LR: positive likelihood ratio; −LR: negative likelihood ratio; CI: confidence interval; MRI: magnetic resonance imaging; SPECT: single-photon emission computerised tomography.

Table 3. Results of all red flags taken during a physical examination in primary care

Study	Index test	Reference standard	Sample size	Prevalence	+LR (95% CI)	−LR (95% CI)
Henschke 2009	Altered sensation (from trunk down)	Follow-up/ imaging	1178	0.007	3.32 (0.22 to 50.86)	0.96 (0.82 to 1.13)
Scavone 1981	Sensation change	X-ray	871	0.030	2.21 (1.14 to 4.27)	0.83 (0.66 to 1.05)
Scavone 1981	Motor deficit	X-ray	871	0.030	2.19 (1.06 to 4.54)	0.86 (0.70 to 1.06)
Scavone 1981	DTR abnormality	X-ray	871	0.030	1.08 (0.37 to 3.18)	0.99 (0.86 to 1.14)
Scavone 1981	Tenderness	X-ray	871	0.030	0.70 (0.25 to 1.97)	1.11 (0.87 to 1.41)
Scavone 1981	Spasm	X-ray	871	0.030	1.25 (0.42 to 3.70)	0.98 (0.85 to 1.12)
Scavone 1981	Sciatica	X-ray	871	0.030	0.42 (0.06 to 2.91)	1.06 (0.98 to 1.15)
Scavone 1981	Hip/leg pain	X-ray	871	0.030	0.21 (0.01 to 3.35)	1.08 (1.02 to 1.14)
Enthoven 2016	Severe disability	X-ray	669	0.045	2.28 (1.26 to 4.14)	0.83 (0.67 to 1.02)
Enthoven 2016	Percussion tenderness of spine	X-ray	669	0.045	1.12 (0.57 to 2.21)	0.97 (0.81 to 1.16)
Enthoven 2016	Sudden decrease in height	X-ray	669	0.045	2.89 (0.90 to 9.24)	0.94 (0.84 to 1.05)
Therriault 2020	Radicular symptoms	X-ray, MRI, or SPECT	1025	0.22	1.91 (0.71 to 5.10)	0.99 (0.97 to 1.01)
Therriault 2020	Numbness or tingling	X-ray, MRI, or SPECT	1025	0.22	0.69 (0.47 to 1.01)	1.07 (1.01 to 1.13)
Therriault 2020	Lumbar extension	X-ray, MRI, or SPECT	1025	0.22	1.27 (1.19 to 1.36)	0.40 (0.28 to 0.57)
Therriault 2020	Lumbar flexion	X-ray, MRI, or SPECT	1025	0.22	1.12 (0.88 to 1.41)	0.96 (0.87 to 1.05)
Therriault 2020	Lumbar rotation	X-ray, MRI, or SPECT	1025	0.22	0.96 (0.66 to 1.41)	1.01 (0.95 to 1.07)
Therriault 2020	Lumbar lateral flexion	X-ray, MRI, or SPECT	1025	0.22	1.15 (0.85 to 1.55)	0.97 (0.90 to 1.04)
Therriault 2020	Single-leg hyperextension test	X-ray, MRI, or SPECT	1025	0.22	1.32 (1.25 to 1.40)	0.19 (0.11 to 0.33)
Therriault 2020	Vertebral tenderness	X-ray, MRI, or SPECT	1025	0.22	0.84 (0.58 to 1.23)	1.03 (0.97 to 1.09)

+LR: positive likelihood ratio; −LR: negative likelihood ratio; CI: confidence interval; DTR: deep tendon reflex; MRI: magnetic resonance imaging; SPECT: single-photon emission computerised tomography.

Table 4. Results of all red flags in secondary care

Study	Index test	Reference standard	Sample size	Prevalence	+LR (95% CI)	−LR (95% CI)
Roman 2010	Age > 52 years	X-ray or CT	1448	0.026	1.52 (1.42 to 1.68)	0.14 (0.04 to 0.53)
Premkumar 2018	Age > 50 years	Imaging[a]	9940	0.056	1.10 (1.05 to 1.16)	0.79 (0.69 to 0.91)
Premkumar 2018	Age > 70 years	Imaging[a]	9940	0.056	1.55 (1.36 to 1.76)	0.86 (0.82 to 0.91)
Kilic 2021	Age 65–75 years	X-ray	136	0.544	0.60 (0.46 to 0.79)	2.51 (1.48 to 4.27)
Kilic 2021	Age ≥ 75 years	X-ray	136	0.544	2.51 (1.48 to 4.27)	0.60 (0.46 to 0.79)
Roman 2010	Female gender	X-ray or CT	1448	0.026	1.51 (1.35 to 1.71)	0.26 (0.10 to 0.65)
Sugiura 2021	Female gender	MRI	101	0.525	0.53 (0.27 to 1.05)	1.26 (0.98 to 1.61)
Roman 2010	Concomitant osteoarthritis	X-ray or CT	1448	0.026	1.05 (0.76 to 1.45)	0.95 (0.69 to 1.32)
Roman 2010	No regular exercise	X-ray or CT	1448	0.026	1.47 (1.25 to 1.75)	0.42 (0.21 to 0.81)
Roman 2010	Absence of leg or buttock pain	X-ray or CT	1448	0.026	2.24 (1.38 to 3.64)	0.80 (0.64 to 0.99)
Roman 2010	BMI < 23	X-ray or CT	1448	0.026	2.22 (1.44 to 3.42)	0.76 (0.59 to 0.97)
Roman 2010	Decreased pain on sitting	X-ray or CT	1448	0.026	1.56 (0.94 to 2.59)	0.87 (0.71 to 1.07)
Roman 2010	No gait abnormality	X-ray or CT	1448	0.026	0.85 (0.68 to 1.08)	1.49 (0.95 to 2.34)
Jin 2020	BPIT	MRI	510	0.68	3.09 (2.47 to 3.85)	0.01 (0.00 to 0.04)
Premkumar 2018	Trauma	Imaging[a]	9940	0.056	2.18 (1.86 to 2.54)	0.85 (0.81 to 0.89)
Sugiura 2021	Participation in sports activities	MRI	101	0.525	1.05 (0.97 to 1.12)	0.18 (0.01 to 3.61)
Sugiura 2021	Lumbar hyperextension	MRI	101	0.525	0.86 (0.73 to 1.02)	2.49 (0.85 to 7.28)
Sugiura 2021	Lumbar hyperflexion	MRI	101	0.525	0.62 (0.45 to 0.86)	2.40 (1.30 to 4.42)
Sugiura 2021	Pain quality	MRI	101	0.525	1.38 (0.97 to 1.96)	0.58 (0.32 to 1.04)
Sugiura 2021	Pain extent	MRI	101	0.525	2.05 (1.22 to 3.44)	0.59 (0.40 to 0.85)
Sugiura 2021	Pain location	MRI	101	0.525	2.40 (1.63 to 3.54)	0.18 (0.08 to 0.39)

[a] Imaging type not reported.
+LR: positive likelihood ratio; −LR: negative likelihood ratio; BMI: body mass index; BPIT: Back Pain-Inducing Test; CI: confidence interval; CT: computed tomography; MRI: magnetic resonance imaging.

Table 5. Results of red flags in tertiary care

Study	Index test	Reference standard	Sample size	Prevalence	+LR	−LR
Gibson 1992	History of direct trauma	X-ray	225	0.06	1.93 (1.48 to 2.52)	0.12 (0.01 to 1.79)
Patrick 1983	Presenting history of trauma	X-ray	552	0.07	1.77 (1.48 to 2.13)	0.36 (0.20 to 0.68)
Reinus 1998	History of trauma (fall)	X-ray	482	0.11	0.18 (0.07 to 0.48)	1.54 (1.38 to 1.71)
Patrick 1983	Sensory deficit	X-ray	552	0.07	1.42 (0.19 to 10.95)	0.99 (0.94 to 1.04)
Patrick 1983	Motor deficit	X-ray	552	0.07	1.39 (0.08 to 25.38)	0.98 (0.96 to 1.03)
Patrick 1983	DTR abnormality	X-ray	552	0.07	1.54 (0.49 to 4.87)	0.97 (0.89 to 1.06)
Gibson 1992	Neurological signs or SLR < 40°, or both	X-ray	225	0.06	2.40 (0.67 to 8.70)	0.81 (0.51 to 1.30)
Reinus 1998	Any motor or sensory dysfunction[a]	X-ray	482	0.11	0.69 (0.22 to 2.17)	1.03 (0.96 to 1.10)
Patrick 1983	Positive SLR	X-ray	552	0.07	1.02 (0.51 to 2.05)	1.00 (0.86 to 1.16)
Patrick 1983	Tenderness	X-ray	552	0.07	1.76 (1.42 to 2.19)	0.47 (0.28 to 0.78)
Patrick 1983	Spasm	X-ray	552	0.07	1.47 (0.83 to 2.60)	0.90 (0.75 to 1.09)
Patrick 1983	Contusion/abrasion	X-ray	552	0.07	31.09 (18.25 to 52.96)	0.15 (0.07 to 0.32)

[a] Weakness, numbness, or paraesthesia present on testing or questioning
+LR: positive likelihood ratio; −LR: negative likelihood ratio; CI: confidence interval; SLR: straight leg raise.

Table 6. Results of combined red flags in primary, secondary, and tertiary care

Study	Index test	Reference standard	Sample size	Prevalence	+LR	−LR
Premkumar 2018	Age > 50 years and trauma	Imaging[a]	9940	0.056	2.54 (2.05 to 3.16)	0.90 (0.87 to 0.94)
Premkumar 2018	Age > 70 years and trauma	Imaging[a]	9940	0.056	4.35 (2.92 to 6.48)	0.96 (0.94 to 0.98)
Gibson 1992	Trauma and neurological signs	X-ray	225	0.06	14.43 (2.38 to 87.64)	0.73 (0.46 to 1.15)
Patrick 1983	Multiple findings	X-ray	552	0.07	2.01 (1.35 to 2.99)	0.73 (0.56 to 0.96)

Table 6. Results of combined red flags in primary, secondary, and tertiary care *(Continued)*

Henschke 2009	1 of 4 red flags +ve	Follow-up/ imaging	1178	0.007	1.75 (1.34 to 2.29)	0.25 (0.04 to 1.57)
Henschke 2009	2 of 4 red flags +ve	Follow-up/ imaging	1178	0.007	15.48 (8.45 to 28.36)	0.39 (0.16 to 0.96)
Henschke 2009	3 of 4 red flags +ve	Follow-up/ imaging	1178	0.007	906.11 (50.37 to 16,299.11)	0.61 (0.36 to 1.03)
Roman 2010	1 of 5 red flags +ve	X-ray or CT	1448	0.026	1.04 (0.98 to 1.10)	0.43 (0.06 to 2.98)
Roman 2010	2 of 5 red flags +ve	X-ray or CT	1448	0.026	1.43 (1.32 to 1.56)	0.16 (0.04 to 0.60)
Roman 2010	3 of 5 red flags +ve	X-ray or CT	1448	0.026	2.45 (2.02 to 2.97)	0.34 (0.19 to 0.61)
Roman 2010	4 of 5 red flags +ve	X-ray or CT	1448	0.026	9.62 (5.88 to 15.73)	0.66 (0.52 to 0.84)
Roman 2010	5 of 5 red flags +ve	X-ray or CT	1448	0.026	7.63 (0.91 to 63.67)	0.98 (0.93 to 1.03)

[a]Imaging type not reported.
+LR: positive likelihood ratio; −LR: negative likelihood ratio; CT: computed tomography.

APPENDICES

Appendix 1. MEDLINE search strategy

I. Index test: clinical red flags

1. exp Medical History Taking/

2. exp Physical Examination/

3. exp "Wounds and Injuries"/

4. exp Accidental Falls/

5. exp Accidents/

6. physical examination*.tw,kf

7. history.tw,kf

8. red flag*.tw,kf

9. physical test*.tw,kf

10. trauma*.tw,kf

11. injur*.tw,kf

12. (clinical* adj (diagnosis or sign* or significance or symptom* or parameter* or assessment* or finding* or evaluat* or indication* or examination*)).tw,kf

13. OR/1–12

II. Population: low back pain

14. exp Back Pain/

15. exp Low Back Pain/

16. exp Sciatica/

17. exp Sciatic Neuropathy/

18. dorsalgia*.tw,kf

19. backache*.tw,kf

20. lumbago.tw,kf

21. back disorder*.tw,kf

22. (lumb* adj pain).tw,kf

23. coccyx.tw,kf

24. coccydynia.tw,kf

25. sciatica.tw,kf

26. spondylosis.tw,kf

27. OR/14–26

III. Target condition: vertebral fracture

28. Fractures, Bone/

29. Fractures, Stress/

30. Fractures, Spontaneous/

31. Fractures, Compression/

32. Fractures, Closed/

33. Spinal Injuries/

34. Spinal Diseases/

35. fracture*.tw,kf

36. OR/28–35

IV. Limits and search combination

37. (20120401-20220726).ed

38. (2012-2022).yr

39. OR/38–39

40. 13 AND 27 AND 36 AND 40

Appendix 2. Embase search strategy

1. exp Anamnesis/

2. exp Physical Examination/

3. exp Physical Performance/

4. exp Accident/

5. physical examination*.tw,kw

6. history.tw,kw

7. red flag*.tw,kw

8. physical test*.tw,kw

9. trauma*.tw,kw

10. injur*.tw,kw

11. (clinical* adj (diagnosis or sign* or significance or symptom* or parameter* or assessment* or finding* or evaluat* or indication* or examination*)).tw,kw

12. OR/1–11

II. Population: low back pain

13. exp Low back Pain/

14. exp Backache/

15. exp Spondylosis/

16. exp Sciatica/

17. exp Ischialgia/

18. dorsalgia.tw,kw

19. back pain.tw,kw

20. backache.tw,kw

21. lumbago.tw,kw

22. back disorder*.tw,kw

23. (lumb* adj pain).tw,kw

24. spondylosis.tw,kw

25. coccyx.tw,kw

26. coccydynia.tw,kw

27. sciatica.tw,kw

28. OR/13–27

III. Target condition: vertebral fracture

29. exp Fracture/

30. exp Spine Fracture/

31. exp Bone Injury/

32. exp Spine Injury/

33. exp Spine Disease/

34. fracture*.tw,kw

35. OR/29–34

IV. Limits and search combination

36. (20120401-20220726).dd

37. (2012-2022).yr

38. OR/36–37

39. 12 AND 28 AND 35 AND 38

40. limit 39 to embase

Appendix 3. CINAHL search strategy

I. Index test: clinical red flags

1. (MH "Patient History Taking")

2. (MH "Diagnostic Tests, Routine")

3. (MM "Physical Examination")

4. (TX "history" or "red flag*" or "physical examination" or "physical test" or "clinical*" or "diagnosis" or TX "sign" or TX "signs" or TX "significance" or TX "symptom*" or TX "parameter*" or TX "assessment" or TX "finding*" or TX "evaluat*" or TX "indication*" or TX "examination*")

5. OR/1–4

II. Population: low back pain

6. (MH "Back Pain+")

7. (MH "Low Back Pain")

8. (MH "Coccyx")

9. (MH "Sciatica")

10. (MH "Lumbar Vertebrae")

11. (MH "Spondylolisthesis")

12. (MH "Spondylolysis")

13. (TX "back pain")

14. (TX "backache")

15. (TX "dorsalgia")

16. (TX "lumbago")

17. (TX "sciatica")

18. (TX "coccyx")

19. (TX "coccydynia")

20. (TX "back disorder*")

21. (TX "lumbago")

22. lumb* w1 pain

23. lumb* n5 pain

24. lumb* n2 vertebra*

25. OR/11–29

III. Target condition: vertebral fracture

26. (MH "Fractures, Closed") or (MH "Fractures, Compression+") or (MH "Fractures, Spontaneous") or (MH "Fractures, Stress+")

27. (MH "Spinal Fractures+")

28. (MH "Spinal Injuries")

29. (MH "Spinal Diseases")

30. (MH "Wounds and Injuries")

31. (TX "fracture*")

32. (TX "trauma")

33. (TX "injury")

34. OR/26–33

IV. Limits and search combination

35. EM 201204-

36. DT 2012-

37. OR/35-36

38. 5 AND 25 AND 34 AND 37

Appendix 4. QUADAS-2 tool: risk of bias and applicability judgements

DOMAIN 1: PATIENT SELECTION	
A. Risk of bias	
Describe methods of patient selection	
• **Was a consecutive or random sample of patients enrolled?** **Score 'yes'** if a consecutive series of patients or a random sample has been selected. Details regarding setting, in inclusion and exclusion criteria, and preferably number of patients eligible and excluded should be given. If a mixed population (e.g. primary and secondary care) is used, the number of participants from each setting will be presented. **Score 'no'** if healthy controls are used. If non-response is high and selective, or there is clear evidence of selective sampling, this will be scored 'no'. **Score 'unclear'** if insufficient information is given on the setting, selection criteria, or selection procedure to make a judgement.	Yes/No/Unclear
• **Did the study avoid inappropriate exclusions?** **Score 'yes'** if participants were excluded based on the prespecified exclusion criteria or if the exclusion criteria was clearly appropriate for the study population. **Score 'no'** if inappropriate exclusion criteria are used such as excluding based on gender, or inappropriate selection bias occurs. **Score 'unclear'** if insufficient information is provided to assess this item.	Yes/No/Unclear
• **Could the selection of patients have introduced bias?** **Score 'low'** if all questions are rated as 'yes'. **Score 'high'** if any of the questions is rated as 'no'. **Score 'unclear** if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is sufficient information to make a decision about the risk of bias.	Risk: Low/High/Unclear

(Continued)

B. Concerns regarding applicability

- **Describe included patients (prior testing, presentation, intended use of index test and setting)**

We expect that most of the diagnostic studies will include participations appropriate for the review question and would have used the index tests described in our inclusion criteria. Many of the index tests are applicable across the different healthcare settings.

- **Is there concern that the included patients do not match the review question?** Concern: Low/High/Unclear

Score 'low concern' if participants meet all inclusion criteria or meet 3 out of the 4 criteria.

Score 'high concern' if participants meet 1 or less of the 4 criteria.

Score 'unclear' if insufficient information is provided to assess this item.

DOMAIN 2: INDEX TEST(S) (if more than 1 index test was used, please complete for each test)

A. Risk of bias

Describe the index test and how it was conducted and interpreted:

- **Were the index test results interpreted without knowledge of the results of the reference standard?** Yes/No/Unclear

Score 'yes' if the results of the index test are interpreted blind to the results of the reference test or if the sequence of testing is always the same (i.e. the index test is always performed first, followed by the reference standard), and consequently the index test is interpreted blind of the reference standard.

Score 'no' if the assessor is aware of the results of the reference standard.

Score 'unclear' if insufficient information is given on independent or blind assessment of the reference standard.

- **If a threshold was used, was it prespecified?** Yes/No/Unclear

Score 'yes' if it is clear that a prespecified threshold was used e.g. the authors follow a study protocol.

Score 'no' if a prespecified threshold was not used or different thresholds were used and not predefined.

Score 'unclear' if insufficient information is provided to assess this item.

- **Could the conduct or interpretation of the index test have introduced bias?** Risk: Low/High/Unclear

Score 'low' if all questions are rated as 'yes'.

Score 'high' if any of the questions is rated as 'no'.

Score 'unclear' if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is enough information to make a decision about the risk of bias.

B. Concerns regarding applicability

- **Is there concern that the index test, its conduct, or interpretation differ from the review question?** Concern: Low/High/Unclear

Score 'low concerns' if the index text performed is the same as specified in the review question and was intended to assess vertebral fracture.

(Continued)

Score 'high concerns' if the index test performed does not follow what is specified in the review question and was not intended to assess vertebral fracture.

Score 'unclear' if insufficient information is provided to assess this item.

DOMAIN 3: REFERENCE STANDARD

A. Risk of bias

Describe the reference standard and how it was conducted and interpreted:

- **Is the reference standard likely to correctly classify the target condition?** — Yes/No/Unclear

Score 'yes' if the following criteria is followed. For traumatic vertebral fractures, CT is preferentially used as the reference test. If CT scan is not available, MRI is used as the reference test. For osteoporotic vertebral fractures and vertebral stress fractures, MRI is preferential used as the reference test. If MRI is not available, bone scintigraphy with SPECT is used as the reference test.

Score 'no' if consensus among observers, or an unknown combination of the clinical assessment ('clinical judgement') is used as reference standard.

Score as 'unclear' if insufficient information is given on the reference standard to make an adequate assessment.

- **Were the reference standard results interpreted without knowledge of the results of the index test?** — Yes/No/Unclear

Score 'yes' if the results of the reference standard are interpreted blind to the results of the index tests or if the sequence of testing is always the same (i.e. the reference standard is always performed first, followed by the index test) and consequently, the reference standard is interpreted blind of the index test.

Score 'no' if the assessor is aware of the results of the index test.

Score 'unclear' if insufficient information is given on independent or blind assessment of the index test.

- **Could the reference standard, its conduct, or its interpretation have introduced bias?** — Risk: Low/High/Unclear

Score 'low' if all questions are rated as 'yes'.

Score 'high' if any of the questions is rated as 'no'.

Score 'unclear' if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is sufficient information to make a decision about the risk of bias.

B. Concerns regarding applicability

- **Is there concern that the target condition as defined by the reference standard does not match the review question?** — Concern: Low/High/Unclear

Score 'low' concern if the target condition defined by the reference standard matches the target condition specified in the review.

Score 'high' concern if the target condition defined by the reference standard does not match the target condition specified in the review.

Score 'unclear' if insufficient information is provided to assess this item.

DOMAIN 4: FLOW AND TIMING

(Continued)

A. Risk of bias	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2×2 table (refer to flow diagram)	
Describe the time interval and any interventions between index test(s) and reference standard:	
• **Was there an appropriate interval between index test(s) and reference standard?** **Score 'yes'** if the time period between clinical assessment and the reference standard ≤ 1 week. **Score 'no'** if the time period between clinical assessment and the reference standard is > 1 week. **Score 'unclear'** if there is insufficient information on the time period between index tests and reference standard.	Yes/No/Unclear
• **Did all patients receive the same reference standard?** **Score 'yes'** if it is clear that all patients receiving the index test are subjected to the same reference standard. **Score 'no'** if different reference standards are used or if only some patients are subjected to the reference standard. **Score 'unclear'** if insufficient information is provided to assess this item.	Yes/No/Unclear
• **Were all patients included in the analysis?** **Score 'yes'** if it is clear that all recruited patients (> 95%) were included in the analysis. **Score 'no'** if the number of patients recruited differs (> 5% difference) from the number of patients included for analysis. **Score 'unclear'** if insufficient information is provided to assess this item.	Yes/No/Unclear
• **Could the patient flow have introduced bias?** **Score 'low'** if all questions are rated as 'yes'. **Score 'high'** if any of the questions is rated as 'no'. **Score 'unclear'** if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is sufficient information to make a decision about the risk of bias.	Risk: Low/High/Unclear

CT: computed tomography; MRI: magnetic resonance imaging; SPECT: single-photon emission computed tomography.

HISTORY

Protocol first published: Issue 7, 2022

CONTRIBUTIONS OF AUTHORS

CSH wrote the subsequent and final (published) versions of the protocol and participated in the conception of the review; design of the review; co-ordination of the review; search and selection of studies for inclusion in the review; collection of data for the review; assessment of the risk of bias in the included studies; analysis of data; interpretation of data; and writing of the review.

MJH provided critical intellectual input and participated in the conception of the review; design of the review; co-ordination of the review; analysis of data; interpretation of data; and writing of the review.

AD provided critical intellectual input and participated in the conception of the review; design of the review; and writing of the review.

JGJ provided critical intellectual input and participated in the conception of the review; design of the review; and writing of the review.

BWK provided critical intellectual input and participated in the conception of the review; design of the review; and writing of the review.

GCM wrote the first draft of the protocol with help from the other authors, provided critical intellectual input, and participated in the conception of the review; design of the review; and writing of the review.

APV provided critical intellectual input and participated in the conception of the review; design of the review; and writing of the review.

CMW provided critical intellectual input and participated in the conception of the review; design of the review; and writing of the review.

QC participated in the conception of the review; design of the review; search and selection of studies for inclusion in the review; collection of data for the review; assessment of the risk of bias in the included studies; and writing of the review.

CGM provided critical intellectual input and participated in the conception of the review; design of the review; co-ordination of the review; analysis of data; interpretation of data, and; writing of the review.

DECLARATIONS OF INTEREST

CSH: none.

MJH: none.

AD: none.

JGJ declares the following: Association of University Radiologists (Fiduciary Officer), Wolters Klewer Health, Inc. (Independent Contractor – Consultant), Springer Publishing (Independent Contractor – Other), GE Healthcare (travel)

BWK: none.

GCM: none.

APV: none.

CMW: none.

QC: none.

CGM is a Senior Editor for Cochrane Musculoskeletal. Editors are required to conduct at least one Cochrane Review, which ensures that they are aware of the processes and commitment needed to conduct reviews.

SOURCES OF SUPPORT

Internal sources

- Vrije Universiteit, EMGO+ Institute for Health and Care Research, Netherlands
- University of Sydney, Australia
- Erasmus MC, University Medical Centre, Netherlands
- The George Institute for Global Health, Australia
- National Institute of Arthritis Musculoskeletal and Skin Diseases (NIAMS)/National Institutes of Health (NIH) P30AR072572, USA

  The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH

External sources

- Dutch Health Care Insurance Board, Netherlands
- National Health and Medical Research Council, Australia

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

In our protocol, we planned to investigate study level covariates in the HSROC analysis to explore heterogeneity but were unable due to a lack of data.

We planned to present results based on the different types of vertebral fracture; however, due to a lack of data this was not feasible. We instead presented our results based on the different healthcare settings and presented the results separately for different vertebral fracture types where available, within each healthcare setting.

We also planned to perform a sensitivity analysis to assess whether diagnostic accuracy differs based on study quality (risk of bias) or the different types of vertebral fracture on study quality but were unable to due to a lack of data. We were unable to perform a sensitivity analysis due to a lack of data.

We planned to use SAS software for meta-analysis; however, meta-analysis was not possible and instead used Review Manager 5 for all statistical analyses and data syntheses (Review Manager 2014).

We intended to present results of how different pretest probability cut-offs and positive or negative results impacted post-test probabilities; however, we were unable to due to a lack of data.

INDEX TERMS

Medical Subject Headings (MeSH)

Adrenal Cortex Hormones; *Contusions; *Fractures, Compression [diagnosis] [diagnostic imaging]; *Fractures, Stress; *Low Back Pain [diagnosis] [etiology]; *Spinal Fractures [diagnosis] [diagnostic imaging]

MeSH check words

Aged; Female; Humans

Chapter Five:

Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review

Statement from co-authors confirming authorship contribution of the PhD candidate

The co-authors of the paper *"Han CS, Hancock MJ, Sharma S, Sharma S, Harris IA, Cohen SP, Magnussen J, Maher CG, Traeger AC. Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review. EClinicalMedicine. 2023 Apr 6;59:101960. doi: 10.1016/j.eclinm.2023.101960. PMID: 37096189; PMCID: PMC10121397."* confirm that Christopher S Han has provided the following contributions to the study:

- Conception and design of the research
- Data acquisition
- Data analysis and interpretation of findings
- Writing of the manuscript and critical appraisal of the content

Christopher S Han _____ Date: 22/11/2023

Professor Chris G Maher _____ Date: 22/11/2023

Professor Mark J Hancock _____ Date: 22/11/2023

Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review

Christopher S. Han,[a,*] Mark J. Hancock,[b] Sweekriti Sharma,[a] Saurab Sharma,[c,d] Ian A. Harris,[a,e] Steven P. Cohen,[f] John Magnussen,[g] Chris G. Maher,[a] and Adrian C. Traeger[a]

[a]The University of Sydney, Sydney Musculoskeletal Health, Sydney, New South Wales, Australia
[b]Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia
[c]Department of Exercise Physiology, School of Health Sciences, Faculty of Medicine and Health, University of New South Wales, Sydney, Australia
[d]Centre for Pain IMPACT, Neuroscience Research Australia, Sydney, Australia
[e]Ingham Institute for Applied Medical Research, School of Clinical Medicine, Faculty of Medicine, UNSW Sydney, Sydney, NSW, Australia
[f]Anaesthesiology, Pain Medicine Division, Johns Hopkins School of Medicine, Baltimore, MD, USA
[g]Department of Clinical Medicine, Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia

eClinicalMedicine 2023;59: 101960

Published Online xxx https://doi.org/10.1016/j.eclinm.2023.101960

Summary

Background The accuracy of diagnostic tests available in primary care to identify the disc, sacroiliac joint, and facet joint as the source of low back pain is uncertain.

Methods Systematic review of diagnostic tests available in primary care. MEDLINE, CINAHL, and EMBASE were searched between March 2006 and 25th January 2023. Pairs of reviewers independently screened all studies, extracted data, and assessed risk of bias using QUADAS-2. Pooling was performed for homogenous studies. Positive likelihood ratios (+LR) ≥2 and negative likelihood ratios (−LR) ≤0.5 were considered informative. This review is registered with PROSPERO (CRD42020169828).

Findings We included 62 studies: 35 investigated the disc, 14 the facet joint, 11 the sacroiliac joint, and 2 investigated all three structures in patients with persistent low back pain. For risk of bias, the domain 'reference standard' scored worst, however approximately half the studies were of low risk of bias for every other domain. For the disc, pooling demonstrated MRI findings of disc degeneration and annular fissure resulted in informative +LRs: 2.53 (95% CI: 1.57–4.07) and 2.88 (95% CI: 2.02–4.10) and −LRs: 0.15 (95% CI: 0.09–0.24) and 0.24 (95% CI: 0.10–0.55) respectively. Pooled results for Modic type 1, Modic type 2, and HIZ on MRI, and centralisation phenomenon yielded informative +LRs: 10.00 (95% CI: 4.20–23.82), 8.03 (95% CI: 3.23–19.97), 3.10 (95% CI: 2.27–4.25), and 3.06 (95% CI: 1.44–6.50) respectively, but uninformative −LRs: 0.84 (95% CI: 0.74–0.96), 0.88 (95% CI: 0.80–0.96), 0.61 (95% CI: 0.48–0.77), and 0.66 (95% CI: 0.52–0.84) respectively. For the facet joint, pooling demonstrated facet joint uptake on SPECT resulted in informative +LRs: 2.80 (95% CI: 1.82–4.31) and −LRs: 0.44 (95% CI: 0.25–0.77). For the sacroiliac joint, a combination of pain provocation tests and absence of midline low back pain resulted in informative +LRs of 2.41 (95% CI: 1.89–3.07) and 2.44 (95% CI: 1.50–3.98) and −LRs of 0.35 (95% CI: 0.12–1.01) and 0.31 (95% CI: 0.21–0.47) respectively. Radionuclide imaging yielded an informative +LR 7.33 (95% CI: 1.42–37.80) but an uninformative −LR 0.74 (95% CI: 0.41–1.34).

Interpretation There are informative diagnostic tests for the disc, sacroiliac joint, and facet joint (only one test). The evidence suggests a diagnosis may be possible for some patients with low back pain, potentially guiding targeted and specific treatment approaches.

Funding There was no funding for this study.

Keywords: Diagnosis; Pathoanatomical; Low back pain; Reference standard; Index test

*Corresponding author.
E-mail address: christopher.han@sydney.edu.au (C.S. Han).

Research in context

Evidence before this study

Our previous review conducted in 2007 found relatively few studies had investigated the diagnostic accuracy of tests to identify the disc, facet joint or sacroiliac joint as the source of low back pain. Pooling was limited due to heterogeneity between studies and poor study quality. The diagnostic accuracy of index tests was unclear.

Added value of this study

In this new study a greater number of studies were able to be pooled, resolving some uncertainty about diagnostic accuracy. Considering an informative diagnostic test as one with a likelihood ratio ≥2.0 or ≤0.5; there were informative tests for the disc, sacroiliac joint, and the facet joint. The review provides evidence that a diagnosis may be possible for some patients with low back pain, potentially guiding clinical management.

Implications of all the available evidence

Our study identifies tests available to primary care clinicians to identify the disc and sacroiliac joint as the source of low back pain and creates opportunities for more targeted and specific treatment approaches. Future research should investigate if this targeted approach provides better outcomes than generic, symptomatic treatment of low back pain.

Introduction

Low back pain is a very common condition, and most people will experience low back pain in their life. For example, the one year incidence of experiencing a first ever episode of low back pain is approximately 6.3%–15.4%, with estimates for the one year incidence of any episode of low back pain being as high as 36%.[1] The one month prevalence of low back pain is estimated to be 23.2% in the general population.[1] The prevalence of low back pain also increases with age,[2] by 80 years old, the prevalence estimates are as high as 40% in males and 35% in females.[2] Low back pain is a symptom and is typically defined as pain between the bottom ribs and the buttock creases.[3]

Currently, most clinical practice guidelines recommend broad treatment approaches such as exercise, medication, or manual therapy for patients with NSLBP[4]; however these interventions typically have small average effects[5] and there is little to no information and guidance provided on how to individualise these treatments for patients with low back pain. Despite the enormous burden of low back pain, the approach to move the field forward remains unclear. One view is that a diagnosis may lead to the development of more specific and targeted treatments.

A diagnosis, however, is controversial in the low back pain field. The dominant view[4] is that for 90% of patients with low back pain (LBP) the pain cannot be attributed to a specific pathology or structure; and these patients should be classified as having non-specific low back pain. An alternate position is that subclassification of non-specific low back pain is possible, by using features from the clinical presentation, clinical assessment, and diagnostic imaging to identify a nociceptive source.[6] The latter approach would allow more precise approaches to the management of low back pain that target an underlying pathology.

The disc, facet joint, and sacroiliac joint (SIJ) are potential sources of low back pain.[7] Most reference standards used to identify the source of low back pain, such as invasive disc provocation procedures or diagnostic blocks, are neither suitable nor available for routine use in primary care. Therefore, there is interest in identifying simpler diagnostic tests to help identify the nociceptive source of low back pain. Our previous 2007 systematic review[7] identified relatively few studies that had investigated the diagnostic accuracy of tests to identify the source of low back pain, limiting the ability to pool results. The systematic review did find that some tests such as MRI evidence of a high intensity zone and disc degeneration, and centralisation phenomenon (i.e., repeated end-range movements that result in distal pain originating from the spine progressively moving towards a central location)[8] increased (when present) or decreased (when absent) the likelihood of the disc being the source of low back pain.[7] A combination of sacroiliac joint provocation tests also modestly increased (when positive) or decreased (when negative) the likelihood of the sacroiliac joint being the source of LBP.[7] No index tests for the facet joint were found to be informative.[7]

The previous review has been highly cited, however it is now over 15 years old.[7] Additional primary studies investigating the diagnostic tests for the disc, facet joint, and sacroiliac joint have since been published and are likely to enable more pooling of results and greater confidence in the findings. Therefore, an update of this review was performed to determine the accuracy of diagnostic tests for the disc, sacroiliac joint, or facet joint as the source of low back pain.

Methods

The systematic review protocol was pre-specified and registered on PROSPERO (CRD42020169828): https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020169828. The current review protocol was registered with PROSPERO (CRD42015029729). Its reporting followed the guidelines of the Preferred Reporting Items of Systematic Reviews and Meta-analyses (PRISMA).[9]

Search strategy

MEDLINE, CINAHL, and EMBASE were searched for the period between March 2006 and 25th January 2023 with the search strategy used in the Hancock et al., 2007 review.[7] The complete search strategies from all databases are presented in Supplementary Appendix S1. We also screened reference lists of included studies. Forward citation searching was also performed. We included studies in all languages.

Eligibility criteria

Studies were required to meet the following criteria (also used in the previous review[7]) to be eligible.

i) Included participants with low back pain without serious pathology such as cancer, infection, or fracture
ii) used a reference standard test advocated by the International Association for the Study of Pain.[10] These were: discography for discogenic pain (with a minimum of two levels tested per patient), intra-articular local anaesthetic blocks for SIJ pain, and either intra-articular blocks or medial branch blocks for facet joint pain
iii) Assessed at least one index test available to primary care clinicians
iv) Presented a 2 × 2 contingency table, or data allowing development of contingency table(s).

One review author (SwS) excluded clearly irrelevant titles. Two review authors (SwS, AT, or SaS) independently screened abstracts to exclude irrelevant studies. Two reviewers (SwS, AT, SaS, or CSH) then independently reviewed full texts for eligibility based on the inclusion criteria. Studies in languages other than English were screened by a researcher fluent in the appropriate language. Disagreements were resolved through discussion and a third reviewer (CGM or MJH).

Data and statistical analysis

Two review authors independently extracted data into the data extraction form. Any disagreements or discrepancies in the data extraction were resolved through discussion and a third reviewer (CGM or MJH) if necessary. This form was used to record study population, hypothesised nociceptive/tissue source of low back pain, index tests, sensitivity, specificity, the positive (+LR) and negative (–LR) likelihood ratios and 95% confidence intervals (95% CI). We prioritised using raw data when available to calculate the sensitivity, specificity, +LR, –LR, and 95% CIs.

Two review authors (SwS, AT, SaS, or CH) used the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) scale[11] independently to rate risk of bias. This tool evaluates four domains such as patient selection; index test; reference standard; and flow and timing.[11] Each author independently rated the risk of bias in each of the four domains (guided by signalling questions). Further details on the decision-making process of study quality are presented in Supplementary Appendix S5. Any disagreements were resolved through discussion and a third reviewer (CGM or MJH).

Following data extraction, data were pooled whenever possible using random effects models. The software used for meta-analysis and to calculate sensitivities, specificities, and likelihood ratios from raw data (2 × 2 tables) was Meta-DiSc v1.4. Due to limited data for most index tests, pooling was not always possible, and results of individual tests were also presented descriptively.

Potential publication bias was assessed post-hoc using the midas command[12] in Stata version 17/BE. To evaluate potential publication bias, funnel plots were explored, and Deeks Funnel Plot Asymmetry Test was used to explore funnel plot asymmetry. Publication bias was deemed statistically significant if the p value was <0.05. As per the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.0, 2022 (Chapter 11.5.5) we only assessed for publication bias if ten or more studies were included in the analysis for a single test.[13]

For index tests scored as either positive or negative we calculated the sensitivity, specificity, and likelihood ratios. We also presented post-test probabilities using standard methods with the "95% confidence interval for post-test probability determined with point estimate of pre-test probability and the 95% confidence interval of the likelihood ratio."[14] For index tests with different thresholds to score as positive or negative, these results are presented descriptively to show the estimates of sensitivities and specificities for each test cut-off.

We focused on likelihood ratios in our results; however, sensitivity and specificity are presented in Table 1 and Supplementary Appendix S6. We considered the index test to be informative if the positive likelihood ratios (+LR) were ≥ 2 and negative likelihood ratios (–LR) were ≤ 0.5.[49,50]

We pre-specified that we would examine the effect of employing stricter reference standards.[51] The stricter reference standards were:[51]

i) Discogenic pain studies: discography with a concordant pain provocation score of 6 out of 10 or greater and an adjacent pain-free control disc;
ii) Facet joint pain studies: greater than or equal to 80% pain relief with double blocks (using a placebo control or comparator local anaesthetic) to account for concurrent pain generators;
iii) Sacroiliac joint pain studies: greater than or equal to 50% pain relief with double blocks (using a placebo control or comparator local anaesthetic) to account for concurrent pain generators.

Index test	Studies/total sample	Sensitivity (95% CI)	Specificity (95% CI)	+LR (95% CI)	−LR (95% CI)
Disc					
Disc degeneration (Grade ≥3)[15–18]	4/381	91.0 (85.7–94.7)	61.3 (54.2–68.0)	2.53 (1.57–4.07)	0.15 (0.09–0.24)
Heterogeneity (I^2)		50.1%	82.7%	83.4%	0.0%
[b]Disc degeneration (Grade ≥4)[16,18,19]	3/288	70.7 (60.7–79.4)	66.7 (59.5–73.3)	2.20 (1.61–3.01)	0.37 (0.19–0.73)
Heterogeneity (I^2)		84.7%	87.7%	44.4%	59.9%
HIZ[18–29]	12/1817	50.1 (46.7–53.4)	86.2 (83.9–88.4)	3.10 (2.27–4.25)	0.61 (0.48–0.77)
Heterogeneity (I^2)		95.4%	65.6%	63.6%	92.2%
Annular fissure[22,25,30–32]	5/920	61.2 (56.3–66.0)	73.8 (69.8–77.5)	2.88 (2.02–4.10)	0.24 (0.10–0.55)
Heterogeneity (I^2)		96.0%	94.3%	79.5%	90.3%
Modic type[18,24,25,33]	4/803	12.9 (9.7–16.6)	98.7 (97.0–99.5)	10.00 (4.20–23.82)	0.84 (0.74–0.96)
Heterogeneity (I^2)		91.0%	0.0%	0.0%	89.2%
Modic type[18,25,34]	3/706	12.0 (8.9–15.9)	98.6 (96.7–99.5)	8.03 (3.23–19.97)	0.88 (0.80–0.96)
Heterogeneity (I^2)		77.3%	0.0%	0.0%	65.8%
[a]Centralisation phenomenon[33,35–37]	4/218	41.2 (33.2–49.6)	85.9 (75.6–93.0)	3.06 (1.44–6.50)	0.66 (0.52–0.84)
Heterogeneity (I^2)		79.9%	66.1%	10.8%	57.9%
SIJ					
Radionuclide imaging (bone scan)[38,39]	2/82	22.7 (11.5–37.8)	96.1 (84.4–99.7)	7.33 (1.42–37.8)	0.74 (0.41–1.34)
Heterogeneity (I^2)		81.4%	0.0%	0.0%	80.2%
[a]Gaenslen's test[40–43]	4/213	47.9 (38.7–57.2)	47.9 (37.5–58.4)	0.85 (0.56–1.28)	1.12 (0.77–1.62)
Heterogeneity (I^2)		72.3%	73.3%	46.5%	34.3%
[a]Sacral thrust test[40,41,43]	3/168	57.3 (45.9–68.2)	48.8 (37.9–59.9)	1.13 (0.73–1.75)	0.87 (0.52–1.44)
Heterogeneity (I^2)		42.5%	84.1%	59.7%	47.7%
[a]Thigh thrust test[40–44]	5/415	54.1 (48.1–60.1)	53.7 (44.9–62.3)	1.13 (0.83–1.55)	0.91 (0.67–1.22)
Heterogeneity (I^2)		74.6%	31.0%	42.5%	42.1%
[a]Compression test[41,43]	2/83	48.6 (31.9–65.6)	71.7 (56.5–84.0)	1.79 (1.03–3.11)	0.74 (0.52–1.05)
Heterogeneity (I^2)		56.0%	0.0%	0.0%	0.0%
[a]Patrick's test (FABER test)[40,43,44]	3/319	76.4 (70.2–81.8)	32.3 (23.3–42.5)	1.05 (0.69–1.60)	0.86 (0.30–2.48)
Heterogeneity (I^2)		80.3%	55.0%	80.7%	84.9%
[a]Distraction test[41,43]	2/82	41.7 (25.5–59.2)	80.4 (66.1–90.6)	2.18 (1.08–4.38)	0.73 (0.54–0.99)
Heterogeneity (I^2)		0.0%	0.0%	0.0%	0.0%
[a]Gillet's test[40,42]	2/134	67.5 (56.4–77.3)	45.5 (31.2–60.2)	1.01 (0.80–1.28)	1.08 (0.75–1.55)
Heterogeneity (I^2)		97.5%	89.2%	23.8%	0.0%
[a]Absence of midline LBP[37,45]	2/226	24.6 (14.1–37.8)	34.3 (27.2–42.0)	2.41 (1.89–3.07)	0.35 (0.12–1.01)
Heterogeneity (I^2)		80.2%	0.0%	0.0%	75.6%
[a]3 or more positive SIJ pain provocation tests[37,41,43,46–48]	6/276	80.5 (72.0–87.4)	68.1 (60.4–75.2)	2.44 (1.50–3.98)	0.31 (0.21–0.47)
Heterogeneity (I^2)		0.0%	69.1%	76.6%	0.0%
Facet joint					
Uptake on SPECT	3/121	72.6 (59.1–83.6)	72.3 (59.8–82.7)	2.80 (1.82–4.31)	0.44 (0.25–0.77)
Heterogeneity (I^2)		31.1%	0.0%	0.0%	7.0%

All magnetic resonance imaging (MRI) findings (index tests) are compared to the absence of the respected MRI finding unless indicated. No index tests based on imaging were able to be pooled for Facet joint. [a]All index tests are based on MRI unless indicated. [b]MRI finding index compared to disc degeneration (Grade ≤3), SIJ = Sacroiliac joint, LR = likelihood ratio, HIZ = high intensity zone, FABER = Flexion abduction external rotation, SPECT = single photon emission computed tomography.

Table 1: **Pooled diagnostic values of index tests.**

We also performed sensitivity analyses post-hoc to examine the effect of study quality (risk of bias) on our findings. We conducted separate sensitivity analyses for each of the four risk of bias domains of the QUADAS-2 scale, each time removing studies at high risk of bias for that domain.

Role of the funding source

There was no funding for this study. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

After removing duplicates (n = 1093) our search identified 3562 citations. Of these, 2964 were deemed clearly irrelevant based on screening by title and a further 496 studies were excluded based on title and abstract. An additional 23 potentially relevant studies were found by contacting experts in the field and relevant citations were provided to search for the additional studies. Following full-text review by two independent authors, 21 new studies and the 41 studies from the previous review were included for analysis (total 62 studies)

(Fig. 1). Excluded studies and primary reasons for exclusion are presented in Supplementary Appendix S2.

Six authors[16,47,52–55] were contacted for additional data and/or discrepancies, with one author[16] able to provide additional data.

The results of the quality assessment using the QUADAS-2[11] scale are presented in Supplementary Appendix S3 Fig. 2 (studies reporting on the disc) and Fig. 3(studies reporting on the SIJ and facet). The domain which scored worst was 'reference standard' where only 17 studies (27%) demonstrated low risk of bias. Low risk of bias scores in the remaining domains were as follows: patient selection (26/62 studies; 42%), index test (34/62 studies; 55%), and flow and timing (30/62 studies; 48%). Seventeen of the 62 (27%) studies demonstrated low concerns for applicability across all domains. Low concerns of applicability scores in the remaining domains were as follows: patient selection (33/62 studies; 53%), index test (54/62 studies; 87%), and reference standard (31/62 studies; 50%). Further details regarding scoring of the domains are presented in Supplementary Appendix S4 and S5.

Of the 62 included studies,[15–42,44–48,52–76] 35 studies[15–36,52,54,55,59,61,66–68,72–76] investigated the disc, 14 studies[56–58,60,62–65,70,71] investigated the facet joint and 11 studies[38–42,44,46–48,53] the sacroiliac joint. Two studies[37,45] investigated all three sources. Studies investigated from 1 to 40 index tests. Sample sizes of the 62 included studies ranged from 15 to 736 (median: 55) and they were conducted in 13 countries. Forty-five studies were conducted in tertiary care, 3 studies in secondary care, and 14 studies were not clear on the healthcare setting. No studies were conducted in primary care. All studies included patients with persistent LBP. The estimated

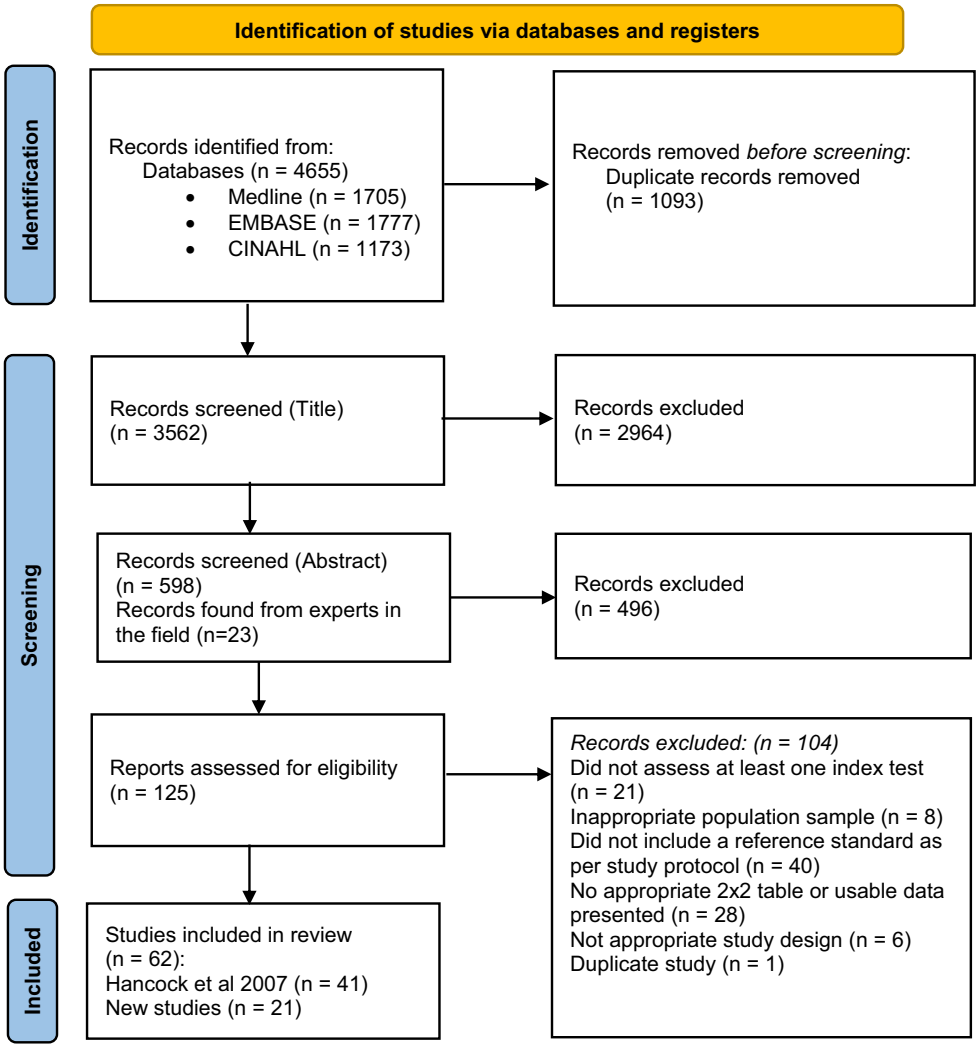


Fig. 1: Study selection.

prevalence of pain originating from discs, facet joints, and the sacroiliac joint (according to a positive reference test) across all studies demonstrated a median of 46% (IQR: 26%) for discogenic pain, 53% (IQR: 7%) for sacroiliac joint pain, and 42% (IQR: 20%) for facet joint pain. The characteristics of the 60 studies are provided in Table 2.

The pooled results are provided in Table 1 and Supplementary Appendix S7. Table 1 and Supplementary Appendix S7 also includes heterogeneity statistics (I^2) for the pooled results. Results for individual studies are provided in Supplementary Appendix S6.

Results for index tests for discogenic pain studies that could be pooled across at least two studies are presented below. All MRI studies investigating discogenic pain calculated diagnostic accuracy at the level of the disc, while studies investigating centralisation always calculated diagnostic accuracy at the level of the patient.

Ten studies[15–19,22,25,52,54,55] investigated MRI evidence of disc degeneration. Five studies[15–19] used the Pfirrmann scale[69] and seven studies measured disc degeneration, four using disc nuclear signal[22,25,68,72] and three using disc height.[19,22,25] Only studies investigating disc degeneration using the Pfirrmann scale (i.e., ≥ grade 3 or ≥ grade 4) were pooled for the main analysis.

For studies measuring disc degeneration using the Pfirrmann scale, we were able to pool four studies investigating disc degeneration with a threshold of ≥ grade 3[15–18] and three studies investigating disc degeneration with a threshold of ≥ grade 4[16,18,19] (Table 1). Pooling demonstrated informative +LRs (2.53; 95% CI: 1.57–4.07, I^2 = 83.4% and 2.20; 95% CI: 1.61–3.01, I^2 = 44.4% for > grade 3 and > grade 4 respectively) and –LRs (0.15; 95% CI: 0.09–0.24, I^2 = 0.0%) and 0.37; 95% CI: 0.19–0.73, I^2 = 59.9% respectively) (Table 1).

Fourteen studies[18–29,52,54] investigated MRI evidence of a high intensity zone (HIZ) and we were able to pool 12 studies[18–29] (Table 1). Pooling demonstrated informative +LRs (3.10; 95% CI: 2.27–4.25, I^2 = 63.6%), but uninformative –LRs (0.61; 95% CI: 0.48–0.77, I^2 = 92.2%) (Table 1).

Six studies[22,25,30–32,74] investigated the MRI evidence of an annular fissure and we were able to pool five studies[22,25,30–32] (Table 1). Pooling demonstrated informative +LRs (2.88; 95% CI: 2.02–4.10, I^2 = 79.5%) and –LRs (0.24; 95% CI: 0.10–0.55, I^2 = 90.3%) (Table 1).

Four studies[15,18,24,25] investigated MRI evidence of Type 1 Modic changes and we were able to pool all 4 studies (Table 1). Pooling demonstrated informative +LRs (10.0; 95% CI: 4.20–23.82, I^2 = 0.0%), but uninformative –LRs (0.84; 95% CI: 0.74–0.96, I^2 = 89.2%) (Table 1).

Three studies[15,18,25] investigated MRI evidence of Type 2 Modic changes and we were able to pool all 3 studies (Table 1). Pooling demonstrated informative +LRs (8.03; 95% CI: 3.23–19.97, I^2 = 0.0%), but uninformative –LRs (0.88; 95% CI: 0.88–0.96, I^2 = 65.8%) (Table 1).

Four studies[33,35–37] investigated the centralisation phenomenon to identify discogenic pain and we were able to pool all four studies (Table 1). Pooling demonstrated informative +LRs (3.06; 95% CI: 1.44–6.50, I^2 = 10.8%), but uninformative –LRs (0.66; 95% CI: 0.52–0.84, I^2 = 57.9%) (Table 1).

Index tests for sacroiliac joint pain studies that could be pooled across at least 2 studies are presented below. Index tests investigated included a range of pain provocation tests (clinical examination) and bone scan.

Two studies[38,39] investigated radionuclide imaging (i.e., bone scan) to identify the sacroiliac joint as a source of low back pain and we were able to pool both studies. Pooling demonstrated informative +LRs (7.33; 95% CI: 1.42–37.8, I^2 = 0.0%), but uninformative –LRs (0.74; 95% CI: 0.41–1.34, I^2 = 80.2%) (Table 1).

For clinical examination-based index tests to identify the sacroiliac joint our meta-analysis demonstrated informative +LR for the distraction test[41,43] (2.18; 95% CI 1.08–4.38, I^2 = 0.0%), but uninformative –LR (0.73; 95% CI: 0.54–0.99, I^2 = 0.0%). Absence of midline low back pain[37,45] demonstrated informative +LR (2.41 95% CI; 1.89–3.07, I^2 = 0.0%) and –LR 0.35 (95% CI: 0.12–1.01, I^2 = 75.6%). Pooling for all other index tests for the sacroiliac joint demonstrated that no test in isolation provided informative +LRs and –LRs (Table 1). Seven studies[37,41–43,46–48] investigated a composite of pain provocation tests (3 or more positive sacroiliac joint provocation tests) and we were able to pool six studies.[37,41,43,46–48] Pooling demonstrated informative +LRs (2.44; 95% CI: 1.50–3.98, I^2 = 76.6%) and –LRs (0.31; 95% CI: 0.21–0.47, I^2 = 0.0%) (Table 1).

Index tests for facet joint pain studies evaluated in two or more individual studies are presented below. Only one index test for the facet joint was able to be pooled. The remaining index tests are presented in Supplementary Appendix S6. Index tests presented below include imaging tests (SPECT—single-photon emission computed tomography), Revel's criteria (5 or more of 7 clinical characteristics: age ≥65, pain relieved by recumbency, pain not exacerbated by coughing, pain not exacerbated by extension/rotation, pain not exacerbated by forward flexion, pain not exacerbated by hyperextension, pain not worse with rising), paraspinal pain, midline spinal pain. Index tests investigated in single studies included a range of clinical examination findings, and clinical prediction rules (Supplementary Appendix S6).

Four studies[56,58,60,77] investigated evidence of facet joint uptake on SPECT and we were able to pool three studies[56,58,77] (Table 1). Pooling demonstrated informative +LRs (2.80; 95% CI: 1.82–4.31, I^2 = 0.0%) and –LRs (0.44; 95% CI: 0.25–0.77, I^2 = 7.0%) (Table 1).

Studies investigating Revel's criteria were not pooled due to heterogeneity in diagnostic values between studies.[63,64,70,71] The diagnostic accuracy of Revel's criteria was inconsistent.[63,64,70,71] Two studies[70,71] found

Study	Country conducted	Tissue source	Sample size	Healthcare setting/Study population	Index tests
April et al., 1992	Australia	Disc	41	Tertiary care: patients with LBP ± leg pain for >3/12 referred for both MRI and discography.	MRI
Akgun et al., 2022	Turkey	Facet joint	51	Tertiary care: patients with LBP, who underwent single-level facet injection	SPECT
Bartynski et al., 2023	USA	Disc	44	Tertiary care: patients with chronic LBP and referred for discography	MRI
Braithwaite et al., 1998	UK	Disc	58	Tertiary care: patients with chronic LBP ± leg pain who underwent MRI and discography prior to spinal fusion	MRI
Carragee et al., 2002	USA	Disc	25	Tertiary care: patients with mild persistent LBP NOT seeking treatment	MRI
Carrera et al., 1980	USA	Facet joint	48	Not clear: patients with buttock pain ± lumbar or lower extremity symptoms referred for diagnostic injections	CT
Chelala et al., 2019*	USA	Disc	736	Tertiary care: patients referred for a discogram were potential surgical candidates or candidates for minimally invasive procedures.	MRI
DePalma et al., 2011	USA	Disc, facet joint, SIJ	160	Tertiary care: patients with LBP that was either recalcitrant to spine-focused physical therapy, oral analgesics, or oral anti-inflammatory medications.	Clinical examination
Donelson et al., 1997	USA	Disc	63	Tertiary care: patients with chronic LBP ± lower extremity pain referred for discography. All patients had one or more positive MRI levels.	Clinical examination
Dreyfuss et al., 1996	USA	SIJ	120	Tertiary care: patients presenting to private pain management centre, with chronic LBP ± lower extremity pain.	Clinical examination
Eskander et al., 2015	USA	SIJ	22	Tertiary care: patients with unilateral LBP ± radiation to lower extremity for >2/12.	Clinical examination, Fluoroscopic exam
Freiermuth et al., 2015	Switzerland	Facet joint	29	Tertiary care: patients with LBP >3/12 referred to pain relief unit.	SPECT/CT
Gornet et al., 2019	USA	Disc	139	Tertiary care: patients with LBP who underwent both MRS and discography.	MRS
Hanna et al., 2013	Sweden	Disc	35	Tertiary care: patients with LBP >6/12 that failed conservative therapy and referred for preoperative lumbar discography	MRI
Helbig et al., 1988	USA	Facet joint	22	Tertiary care: patients who underwent lumbar facet joint injection for LBP and leg pain	Clinical examination, radiograph
Holder et al., 1995	USA	Facet joint	19	Tertiary care: patients referred with a diagnosis of possible facet syndrome	SPECT
Horton et al., 1992	USA	Disc	25	Tertiary care: patients being considered for operative intervention who underwent MRI and discography.	MRI
Ito et al., 1998	USA	Disc	39	Not clear: patients with chronic LBP studied with MRI and discography	MRI
Kang et al., 2009	South Korea/USA	Disc	62	Tertiary care: patients with severe LBP that was likely to be disc related ± radiation to lower limb extremity.	MRI
Koh et al., 2011	South Korea	Facet joint	33	Not clear: patients with chronic LBP >6/12.	SPECT
Kokkonen et al., 2002	Finland	Disc	36	Tertiary care: patients admitted to hospital because of chronic LBP of unclear but suspected discogenic origin.	MRI
Lam et al., 2000	UK	Disc	73	Tertiary care: patients considered for spinal fusion. All patients had at least one positive HIZ	MRI
Laslett et al., 2003	USA	SIJ	200	Tertiary care: patients presenting to private pain management centre, with chronic LBP ± lower extremity pain.	Clinical examination
Laslett et al., 2004	USA	Facet joint	116	Tertiary care: patients with chronic LBP ± lower extremity symptoms referred to diagnostic radiology practice.	Clinical examination
Laslett et al., 2005a	USA	Disc	83	Tertiary care: patients with chronic LBP ± leg pain and abnormal MRI referred to diagnostic injections	Clinical examination
Laslett et al., 2005b	USA	SIJ	32	Tertiary care: patients with unilateral LBP ± posterior thigh pain for more than 7 weeks. All patients had tenderness over SIJ joint line.	Clinical examination
Laslett et al., 2006a	USA	Disc	216	Tertiary care: patients with persistent LBP ± radiation to lower limb extremity referred to a private radiology practice specializing in the diagnosis of spinal pain.	Clinical examination
Laslett et al., 2006b	USA	Facet joint	151	Tertiary care: patients with persistent LBP ± radiation to lower limb extremity referred to a private radiology practice specializing in the diagnosis of spinal pain.	Clinical examination

(Table 2 continues on next page)

Study	Country conducted	Tissue source	Sample size	Healthcare setting/Study population	Index tests
(Continued from previous page)					
Lei et al., 2008	UK	Disc	55	Tertiary care: patients with disabling LBP that were likely to be disc related. Candidates for spinal surgery also included.	MRI
Lewinnek et al., 1986	USA	Facet joint	21	Not clear: patients with LBP	Clinical examination
Lim et al., 2005	South Korea	Disc	47	Not clear: patients with chronic LBP who underwent both discography and MRI. Disc degeneration but no neural compression on MRI.	MRI
Maigne et al., 1998	France	SIJ	42	Tertiary care: patients with LBP >3/12 referred for facet injection. 40 patients with LBP but not true sciatica 86 chronic LBP patients referred for facet joint blocks 10 patients with LBP ± leg pain	Radionuclide image
Manchikanti et al., 1999	USA	Facet joint	120	Secondary care: patients presenting to private pain management centre, with chronic LBP ± lower extremity pain.	Clinical examination
Manchikanti et al., 2000	USA	Facet joint	200	Secondary care: patients presenting to private pain management centre, with chronic LBP ± lower extremity pain.	Clinical examination
Mekhail 2021	USA	SIJ	199	Tertiary care: patients with LBP evaluated by diagnostic SIJ injections	Clinical examination
Milette et al., 1990	Canada	Disc	100	Not clear: patients with subjective complaints suggestive of HNP but no neurological deficits.	CT
Nejati 2020	Iran	SIJ	48	Secondary care: patients with LBP or buttock pain with or without lower extremity pain	Clinical examination
O'Neill et al., 2008	Canada	Disc	143	Tertiary care: patients with LBP >6/12. No neurologic deficits, no previous surgery, and had an MRI.	MRI
Ohnmeiss et al., 1999	USA	Disc	187	Not clear: patients with LBP ± lower extremity pain who underwent discography.	Clinical examination
Osti et al., 1992	Australia	Disc	33	Tertiary care: patients investigated for LBP	MRI
Parker et al., 1996	USA	Disc	15	Tertiary care: patients with chronic LBP ± leg pain referred for MRI and discography prior to surgery.	MRI
Revel et al., 1992	France	Facet joint	60	Not clear: patients with chronic LBP below L5, over one SIJ ± leg pain referred for diagnostic injections.	Clinical examination
Revel et al., 1998	France	Facet joint	42	Tertiary care: patients with LBP >3/12 referred for facet injection. 40 patients with LBP but not true sciatica 86 chronic LBP patients referred for facet joint blocks 10 patients with LBP ± leg pain	Clinical examination
Ricketson et al., 1996	USA	Disc	29	Not clear: patients with LBP ± radicular symptoms who underwent both MRI and discography.	MRI
Saifuddin et al., 1998	UK	Disc	58	Tertiary care: patients with chronic non-radicular LBP who underwent both MRI and discography before possible spinal fusion	MRI
Schellhas et al., 1996	USA	Disc	63	Not clear: patients who underwent MRI and Discography for LBP and had at least one positive HIZ level	MRI
Schneider et al., 2020	USA	SIJ	39	Tertiary care: patients referred for SIJ injection	Clinical examination
Simmons et al., 1991	USA	Disc	164	Not clear: patients with LBP ± radicular symptoms	MRI
Slipman et al., 1996	USA	SIJ	50	Tertiary care: patients referred to spine centre with LBP ± leg pain. Positive physical examination on at least 3 SIJ tests.	Radionuclide image
Smith et al., 1998	USA	Disc	55	Tertiary care: patients who underwent both lumbar discography and MRI.	MRI
Stanford et al., 2010	Canada	SIJ	34	Not clear: patients with LBP that was refractory to non-invasive conservative spine treatment for >6/12	Clinical examination
Stojanovic et al., 2010	USA	Facet joint	119	Not clear: patients with chronic LBP unresponsive to conservative therapy	MRI
van der Wurff et al., 2006	Netherlands	SIJ	85	Tertiary care: patients referred for SIJ blocks, with pain principally below L5	Clinical examination
Vanharanta et al., 1988	USA	Disc	107	Not clear: patients with LBP where "discography was indicated" and plain radiograph had been previously performed.	Plain radiograph
Vanharanta et al., 1998	USA	Disc	78	Tertiary care: patients with chronic LBP ± leg pain, who underwent MRI and discography	MRI, vibration
Waldrop et al., 2021	USA	Disc	736	Tertiary care: patients referred for lumbar spine discography	MRI
Weishaupt et al., 2001	Switzerland	Disc	50	Tertiary care: patients with severe chronic LBP who were candidates for surgery and had abnormal MRI.	MRI

(Table 2 continues on next page)

Study	Country conducted	Tissue source	Sample size	Healthcare setting/Study population	Index tests
(Continued from previous page)					
Yoshida et al., 2002	Japan	Disc	23	Tertiary care: patients with chronic LBP and leg pain who underwent both MRI and discography.	MRI
Young et al., 2003	USA	Disc, SIJ, Facet joint	81	Tertiary care: patients with chronic lumbopelvic pain referred for diagnostic injections.	Clinical examination
Yrjama et al., 1994	Finland	Disc	57	Tertiary care: patients with LBP	Vibration
Yrjama et al., 1996	Finland	Disc	38	Tertiary care: patients with LBP mostly for several years	Ultrasound, vibration
Yrjama et al., 1997	Finland	Disc	33	Tertiary care: patients with chronic LBP mostly for many years	MRI, vibration

SIJ = sacroiliac joint; LBP = low back pain; MRI = magnetic resonance imaging; MRS = magnetic resonance spectroscopy; CT = computed tomography; SPECT = single-photon emission computerized tomography; HIZ = high intensity zone; HNP = herniated nucleus pulposus; USA = United Stated of America; UK = United Kingdom. *Most index tests in this study[52] were excluded as this study only reported the presence or absence of the "worst" disc feature present. For example, if a disc presented with disc degeneration and disc herniation then that disc was categorised as disc herniation only.

Table 2: **Individual study characteristics.**

informative +LRs and –LRs and two other studies[63,64] found uninformative +LRs and –LRs (Supplementary Appendix S6). None of the seven individual items comprising Revels criteria produced informative +LRs and –LRs in more than one study (Supplementary Appendix S6). Para-spinal pain and midline spinal pain both had uninformative +LRs and –LRs (Supplementary Appendix S6).

We also pre-planned sensitivity analysis to investigate the influence of the reference test quality on estimates of the diagnostic accuracy of index tests, by limiting the studies to those meeting our higher threshold for the reference test quality.

Five studies[20,21,25,26,28] investigating HIZ met our criteria for a higher quality reference standard. When pooling only these studies[20,21,25,26,28] the results demonstrated slightly higher informativeness for the +LRs (3.54; 95% CI: 2.03–6.20) compared to when pooling all studies regardless of the reference standard quality (+LR: 3.10; 95% CI: 2.27–4.25) (Table 1). The pooled –LR (0.48; 95% CI: 0.28–0.81) met our cut-off for informativeness unlike when all studies were pooled regardless of reference standard quality (–LR: 0.61; 95% CI: 0.48–0.77) (Supplementary Appendix S8).

Two studies[46,48] investigating a combination of sacroiliac joint pain provocation tests met our criteria for a higher quality reference standard. When pooling only these studies[46,48] the results demonstrated higher informativeness for the +LR (4.09; 95% CI: 2.53–6.60) compared to when pooling all studies regardless of reference standard quality (+LR 2.44; 95% CI: 1.50–3.98) (Table 1). The pooled –LR (0.17; 95% CI: 0.08–0.39) met our cut-off for informativeness unlike when all studies were pooled irrespective of reference standard quality (–LR 0.31; 95% CI: 0.21–0.47) (Supplementary Appendix S8). However, these results should be interpreted with caution due to the small number of studies.

We also performed sensitivity analyses post-hoc to examine the influence of risk of bias quality on estimates of the diagnostic accuracy of index tests by excluding studies with high risk of bias in each of the four domains of the QUADAS-2 scale. For a combination of sacroiliac pain provocation tests, removing studies of high risk of bias for the domain 'patient selection' reduced the +LR from 2.44 (95% CI: 1.50–3.98) to 1.69 (95% CI: 0.98–2.90) and for 'reference standard' increased the +LR from 2.44 (95% CI: 1.50–3.98) to 3.06 (95% CI: 1.75–5.33). For HIZ and the remaining domains for combination of sacroiliac joint pain provocation tests, the removal of studies with high risk of bias had little to no effect on the diagnostic accuracy values. Further details are presented in Supplementary Appendix S8.

Only one index test had more than ten studies in the meta-analysis to explore publication bias. No publication bias was found for the meta-analysis of the index test 'MRI evidence of high intensity zone' (12 studies) based on Deeks test (p = 0.53).

Discussion

We located 60 studies investigating the diagnostic accuracy of tests to identify the source of low back pain. To our knowledge this is the most comprehensive analysis of diagnostic tests for low back pain to date. Most studies focussed on the disc (35 studies) with fewer studies considering the facet joint (14 studies) or the sacroiliac joint (11 studies). We found evidence that some index tests had informative +LRs and –LRs for the disc and sacroiliac joint, and the facet joint. In studies of people with persistent pain, MRI findings of disc degeneration, HIZ, annular fissure, Modic type 1, Modic type 2, and uptake of facet joint on SPECT increased the likelihood of the disc being a nociceptive source of low back pain. The only informative physical examination-based index tests were a positive centralisation phenomenon to identify discogenic pain, and the absence of midline low back pain and a combination of sacroiliac joint pain provocation tests to identify sacroiliac joint pain.

In the previous review, MRI evidence of a high intensity zone and disc degeneration, and centralisation phenomenon were found to be informative to identify the disc as the source of low back pain and a combination of sacroiliac joint provocation tests were found to be informative to identify the SIJ as the source of low back pain.[7] The results of our review reinforced the results of the previous review as a larger number of studies were able to be pooled. The previous review found there was no informative index tests to identify the facet joint as the source of low back pain, however the results of our review found that facet joint uptake on SPECT was informative.

Strengths of this study include using a sensitive search strategy and following a strict pre-specified protocol. Another strength is that we were able to pool a greater number of studies compared to the previous review resulting in more confidence in the results.[7] A limitation of our review is that most studies included in our review included convenience samples of patients referred to tertiary settings for further diagnostic testing, despite testing index tests more commonly used in primary care. This sampling approach may inflate prevalence estimates, but it's impact on diagnostic accuracy of the index tests is unclear. Risk of bias was generally high for the domain 'reference standard' as most studies used a less strict criteria for a positive test, which may reflect the realities of how these tests are used in clinical practice. Approximately half of the studies in the other three domains demonstrated low risk of bias, which should provide clinicians with some confidence in our results. Our review found that 21 new studies investigating the source of low back pain have been published in the past 15 years. While this substantially increased the available evidence, some index tests still lack adequate evidence and demonstrate imprecise values to draw confident conclusions regarding diagnostic accuracy.

Our results challenge the dominant view in the low back pain field, that a pathoanatomical diagnosis is usually not possible and so the label non-specific low back pain should instead be used for most patients. Our review provides preliminary evidence that a diagnosis may be possible for a subgroup of patients with low back pain, potentially moving beyond the non-specific low back pain label. The ability to form a diagnosis in people with persistent low back pain is an important step towards developing new, more targeted, and specific treatment approaches. The pathology and causal mechanism driving Modic type 1 changes have, for example, been linked to bacterial infection, inflammation, and bone marrow oedema.[78] For annular fissures, zones of granulation tissue have been proposed to be the causal mechanism driving discogenic low back pain.[79] Treatments that aim to address these causes and mechanisms (e.g., antibiotics, Zoledronic acid or Denosumab for Modic type 1 changes) require much further research regarding their safety and effectiveness; however, these research approaches present an opportunity to develop targeted more effective treatments.

Our results align with a recent systematic review[80] that investigated the diagnostic accuracy of clusters of pain provocation tests for the sacroiliac joint as the source of LBP. The review found similar LRs to our review for 3 or more sacroiliac joint provocation tests (+LR: 2.13; 95% CI: 1.20–3.90 and –LR: 0.33; 95% CI 0.11–0.72).[80] The review also found a positive or negative test was associated with an increase or decrease in the probability of the sacroiliac joint being the source of pain.[80] Our results challenge and support some of the clinical features that are thought to be associated with discogenic and SIJ pain. For example, a recent seminar in The Lancet[6] advocated the MRI findings that our review showed were informative for discogenic LBP, but it also advocated clinical features which lack evidence of diagnostic accuracy. Our review will thus have value for informing clinical assessment of low back pain.

Studies investigating index tests with different thresholds or defined by different measures (e.g., disc degeneration) did not always make it clear which thresholds or definitions were being compared. It is possible this may have affected diagnostic accuracy values; however, these studies were reviewed independently by two reviewers (CSH, MJH, or CGM) to minimise errors. Possible sources of heterogeneity in our review could be from the population (e.g., were patients' surgical candidates), healthcare setting (e.g., primary care verse secondary care), the index test performed, the reference standard, and methods of the study (e.g., defined thresholds for a positive test). Unfortunately, it was not possible to explore heterogeneity between studies due to a lack of data and studies that could be pooled in the analysis.

There is a need for more diagnostic research evaluating tests to identify the pathoanatomical source of low back pain. To date the diagnostic research has focussed on the facet joint, disc and sacroiliac joint as sources of low back pain, but other structures of the lumbosacral spine (e.g., muscles, fascia, ligaments and vertebral body)[6] are potential nociceptive sources and there are treatments targeting these structures. As there are now informative tests for the disc, sacroiliac joint, and facet joint (only one test), it should be a priority to investigate whether patients judged to have these conditions based upon validated index tests have different prognoses or responses to treatment compared to patients considered not to have those conditions. Ultimately, the diagnoses based upon these index tests will only have utility if they predict prognosis or response to treatment. A related issue is how informative does a test need to be to justify shifting from generic treatment of cases of non-specific low back pain to more targeted treatments directed towards the structure putatively responsible for a patient's symptoms. This issue is of interest as the +LRs for the

disc tests ranged from 2.20 to 10.00 and so yield quite different certainties in a particular diagnosis. Some of the index tests investigated in our review are continuous (e.g., disc degeneration) and the diagnostic accuracy may be influenced by the thresholds used to define a positive test. Future research should more carefully investigate the optimal thresholds for a positive test.

Not all patients with low back pain would potentially benefit from a pathoanatomical diagnosis as many acute cases settle rapidly with little or no formal treatment.[81] The potential benefit is more likely to be seen in those patients with persisting low back pain where a pathoanatomical diagnosis may help select more targeted treatment. However even in this patient group it would be important to consider issues such as feasibility, access, and costs of the index tests, and of the resulting treatment approaches they would lead to.

A better understanding of the clinical value of these diagnostic tests will allow clinicians to be selective in which patients are most likely to benefit from further testing. Future research into the utility of these diagnostic tests is important as uncritical application may lead to unnecessary imaging, overdiagnosis, and ultimately, overtreatment. This is important as research has shown unnecessary testing can lead to increases in resources, costs, and negative downstream consequences.[82] Some of the informative index tests included in this review require imaging (e.g., MRI or SPECT) which is accessible in some but not all countries. These imaging tests are reasonably expensive tests that may restrict access for many people, but if they inform management that aids recovery and return to work, they may be cost-effective. Other index tests include clinical examination tests such as those for SIJ pain, which are cheap but do require adequate clinical skills and training.

In conclusion, our review identified informative index tests to identify the disc, sacroiliac joint, and the facet joint as the source of LBP. This suggests it may be possible to diagnose a subgroup of patients with low back pain. A diagnosis may shift generic symptomatic treatments of non-specific low back pain to instead, more specific treatments targeted to the pathoanatomical source of low back pain. Our results challenge and support some of the diagnostic tests thought to be associated with discogenic and SIJ pain. Further high-quality primary studies are required to assess the clinical utility of these tests in guiding more targeted treatments and subsequently improving outcomes, ultimately to improve patient care.

Contributors

Conception of the study, drafting study protocol: All authors.
Acquisition of data: CSH, AT, SwS, SaS, SPC, CGM, MJH.
Analysis and/or interpretation of data: CSH, AT, JM, IAH, CGM, MJH.
Drafting the manuscript: CSH, AT, CGM, MJH.
Final approval of the manuscript: All authors.

Data sharing statement

The original data will be shared with researchers upon request.

Declaration of interests

We (author team) declare no competing interests.

Acknowledgements

CGM is funded by an NHMRC Research Fellowship (APP 1194283).
No other funding was provided to complete this study.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.101960.

References

1 Hoy D, Brooks P, Blyth F, Buchbinder R. The Epidemiology of low back pain. *Best Pract Res Clin Rheumatol*. 2010;24(6):769–781.
2 Hoy D, March L, Brooks P, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Ann Rheum Dis*. 2014;73(6):968–974.
3 Dionne CE, Dunn KM, Croft PR, et al. A consensus approach toward the standardization of back pain definitions for use in prevalence studies. *Spine (Phila Pa 1976)*. 2008;33(1):95–103.
4 Oliveira CB, Maher CG, Pinto RZ, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *Eur Spine J*. 2018;27(11):2791–2803.
5 Koes BW, van Tulder M, Lin CW, Macedo LG, McAuley J, Maher C. An updated overview of clinical guidelines for the management of non-specific low back pain in primary care. *Eur Spine J*. 2010;19(12):2075–2094.
6 Knezevic NN, Candido KD, Vlaeyen JWS, Van Zundert J, Cohen SP. Low back pain. *Lancet*. 2021;398(10294):78–92.
7 Hancock MJ, Maher CG, Latimer J, et al. Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. *Eur Spine J*. 2007;16(10):1539–1550.
8 Apeldoorn AT, van Helvoirt H, Meihuizen H, et al. The influence of centralization and directional preference on spinal control in patients with nonspecific low back pain. *J Orthop Sports Phys Ther*. 2016;46(4):258–269.
9 Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;151(4):264–269. W264.
10 International Organization for the Study of Pain. Descriptions of chronic pain syndromes and definitions of pain terms. *Pain*. 2011.
11 Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536.
12 Dwamena B, Sylvester R, Carlos R. *midas: Stata module for meta-analytical integration of diagnostic test accuracy studies*. Statistical Software Components S456880, Department of Economics, Boston College. 2007.
13 Deeks JJ, B P, Leeflang MM, Takwoingi Y. Chapter 11: presenting findings. Draft version (4 October 2022) for inclusion. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, eds. *Cochrane Handbook for systematic reviews of diagnostic test accuracy version 2 London*. Cochrane; 2022.
14 Downie A, Williams CM, Henschke N, et al. Red flags to screen for malignancy and fracture in patients with low back pain: systematic review. *BMJ*. 2013;347:f7095.
15 Braithwaite I, White J, Saifuddin A, Renton P, Taylor BA. Vertebral end-plate (Modic) changes on lumbar spine MRI: correlation with pain reproduction at lumbar discography. *Eur Spine J*. 1998;7(5):363–368.
16 Carragee EJ, Alamin TF, Miller J, Grafe M. Provocative discography in volunteer subjects with mild persistent low back pain. *Spine J*. 2002;2(1):25–34.
17 Parker LM, Murrell SE, Boden SD, Horton WC. The outcome of posterolateral fusion in highly selected patients with discogenic low back pain. *Spine (Phila Pa 1976)*. 1996;21(16):1909–1916. discussion 1916-1907.
18 Weishaupt D, Zanetti M, Hodler J, et al. Painful lumbar disk derangement: relevance of endplate abnormalities at MR imaging. *Radiology*. 2001;218(2):420–427.
19 Lim CH, Jee WH, Son BC, Kim DH, Ha KY, Park CK. Discogenic lumbar pain: association with MR imaging and CT discography. *Eur J Radiol*. 2005;54(3):431–437.

20 Aprill C, Bogduk N. High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging. *Br J Radiol*. 1992;65(773):361–369.
21 Hanna H, Tommy H, Hebelka H, Hansson T. HIZ's relation to axial load and low back pain: investigated with axial loaded MRI and pressure controlled discography. *Eur Spine J*. 2013;22(4):734–739.
22 Ito M, Incorvaia KM, Yu SF, Fredrickson BE, Yuan HA, Rosenbaum AE. Predictive signs of discogenic lumbar pain on magnetic resonance imaging with discography correlation. *Spine (Phila Pa 1976)*. 1998;23(11):1252–1258. discussion 1259-1260.
23 Lam KS, Carlin D, Mulholland RC. Lumbar disc high-intensity zone: the value and significance of provocative discography in the determination of the discogenic pain source. *Eur Spine J*. 2000;9(1):36–41.
24 Lei D, Rege A, Koti M, Smith FW, Wardlaw D. Painful disc lesion: can modern biplanar magnetic resonance imaging replace discography? *J Spinal Disord Tech*. 2008;21(6):430–435.
25 O'Neill C, Kurgansky M, Kaiser J, Lau W. Accuracy of MRI for diagnosis of discogenic pain. *Pain Physician*. 2008;11(3):311–326.
26 Ricketson R, Simmons JW, Hauser BO. The prolapsed intervertebral disc. The high-intensity zone with discography correlation. *Spine (Phila Pa 1976)*. 1996;21(23):2758–2762.
27 Saifuddin A, Braithwaite I, White J, Taylor BA, Renton P. The value of lumbar spine magnetic resonance imaging in the demonstration of anular tears. *Spine (Phila Pa 1976)*. 1998;23(4):453–457.
28 Schellhas KP, Pollei SR, Gundry CR, Heithoff KB. Lumbar disc high-intensity zone. Correlation of magnetic resonance imaging and discography. *Spine (Phila Pa 1976)*. 1996;21(1):79–86.
29 Smith BM, Hurwitz EL, Solsberg D, et al. Interobserver reliability of detecting lumbar intervertebral disc high-intensity zone on magnetic resonance imaging and association of high-intensity zone with pain and anular disruption. *Spine (Phila Pa 1976)*. 1998;23(19):2074–2080.
30 Bartynski WS, Agarwal V, Trang H, et al. Enhancing annular fissures and high-intensity zones: pain, internal derangement, and anesthetic response at provocation lumbar discography. *AJNR Am J Neuroradiol*. 2023;44(1):95–104.
31 Vanharanta H, Ohnmeiss DD, Aprill CN. Vibration pain provocation can improve the specificity of MRI in the diagnosis of symptomatic lumbar disc rupture. *Clin J Pain*. 1998;14(3):239–247.
32 Yoshida H, Fujiwara A, Tamai K, Kobayashi N, Saiki K, Saotome K. Diagnosis of symptomatic disc by magnetic resonance imaging: T2-weighted and gadolinium-DTPA-enhanced T1-weighted magnetic resonance imaging. *J Spinal Disord Tech*. 2002;15(3):193–198.
33 Donelson R, Aprill C, Medcalf R, Grant W. A prospective study of centralization of lumbar and referred pain. A predictor of symptomatic discs and anular competence. *Spine (Phila Pa 1976)*. 1997;22(10):1115–1122.
34 Horton WC, Daftari TK. Which disc as visualized by magnetic resonance imaging is actually a source of pain? A correlation between magnetic resonance imaging and discography. *Spine (Phila Pa 1976)*. 1992;17(6 Suppl):S164–S171.
35 Laslett M, Aprill CN, McDonald B, Oberg B. Clinical predictors of lumbar provocation discography: a study of clinical predictors of lumbar provocation discography. *Eur Spine J*. 2006;15(10):1473–1484.
36 Laslett M, Oberg B, Aprill CN, McDonald B. Centralization as a predictor of provocation discography results in chronic low back pain, and the influence of disability and distress on diagnostic power. *Spine J*. 2005;5(4):370–380.
37 Young S, Aprill C, Laslett M. Correlation of clinical examination characteristics with three sources of chronic low back pain. *Spine J*. 2003;3(6):460–465.
38 Maigne JY, Boulahdour H, Chatellier G. Value of quantitative radionuclide bone scanning in the diagnosis of sacroiliac joint syndrome in 32 patients with low back pain. *Eur Spine J*. 1998;7(4):328–331.
39 Slipman CW, Sterenfeld EB, Chou LH, Herzog R, Vresilovic E. The value of radionuclide imaging in the diagnosis of sacroiliac joint syndrome. *Spine (Phila Pa 1976)*. 1996;21(19):2251–2254.
40 Dreyfuss P, Michaelsen M, Pauza K, McLarty J, Bogduk N. The value of medical history and physical examination in diagnosing sacroiliac joint pain. *Spine (Phila Pa 1976)*. 1996;21(22):2594–2602.
41 Laslett M, Aprill CN, McDonald B, Young SB. Diagnosis of sacroiliac joint pain: validity of individual provocation tests and composites of tests. *Man Ther*. 2005;10(3):207–218.
42 Nejati P, Sartaj E, Imani F, Moeineddin R, Nejati L, Safavi M. Accuracy of the diagnostic tests of sacroiliac joint dysfunction. *J Chiropr Med*. 2020;19(1):28–37.
43 Schneider BJ, Ehsanian R, Rosati R, Huynh L, Levin J, Kennedy DJ. Validity of physical exam maneuvers in the diagnosis of sacroiliac joint pathology. *Pain Med*. 2020;21(2):255–260.
44 Mekhail N, Saweris Y, Sue Mehanny D, Makarova N, Guirguis M, Costandi S. Diagnosis of sacroiliac joint pain: predictive value of three diagnostic clinical tests. *Pain Pract*. 2021;21(2):204–214.
45 Depalma MJ, Ketchum JM, Trussell BS, Saullo TR, Slipman CW. Does the location of low back pain predict its source? *PM R*. 2011;3(1):33–39.
46 Laslett M, Young SB, Aprill CN, McDonald B. Diagnosing painful sacroiliac joints: a validity study of a McKenzie evaluation and sacroiliac provocation tests. *Aust J Physiother*. 2003;49(2):89–97.
47 Stanford G, Burnham RS. Is it useful to repeat sacroiliac joint provocative tests post-block? *Pain Med*. 2010;11(12):1774–1776.
48 van der Wurff P, Buijs EJ, Groen GJ. A multitest regimen of pain provocation tests as an aid to reduce unnecessary minimally invasive sacroiliac joint procedures. *Arch Phys Med Rehabil*. 2006;87(1):10–14.
49 Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ*. 2004;329(7458):168–169.
50 Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The evidence-based medicine working group. *JAMA*. 1994;271(9):703–707.
51 Petersen T, Laslett M, Juhl C. Clinical classification in low back pain: best-evidence diagnostic rules based on systematic reviews. *BMC Musculoskeletal Disord*. 2017;18(1):188.
52 Chelala L, Trent G, Waldrop G, Dagher AP, Reinig JW. Positive predictive values of lumbar spine magnetic resonance imaging findings for provocative discography. *J Comput Assist Tomogr*. 2019;43(4):568–571.
53 Eskander JP, Ripoll JG, Calixto F, et al. Value of examination under fluoroscopy for the assessment of sacroiliac joint dysfunction. *Pain Physician*. 2015;18(5):E781–E786.
54 Kang CH, Kim YH, Lee SH, et al. Can magnetic resonance imaging accurately predict concordant pain provocation during provocative disc injection? *Skeletal Radiol*. 2009;38(9):877–885.
55 Waldrop G, Trent G, Dagher AP, Reinig J, Thompson KJ. The association between magnetic resonance imaging disc pathology and provocative discography at the lumbar level. *J Comput Assist Tomogr*. 2021;45(1):146–150.
56 Akgun E, Akgun MY. The effectiveness of bone scintigraphy in the management of low back pain. *Clin Neurol Neurosurg*. 2022;222:107440.
57 Carrera GF. Lumbar facet joint injection in low back pain and sciatica: preliminary results. *Radiology*. 1980;137(3):665–667.
58 Freiermuth D, Kretzschmar M, Bilecen D, et al. Correlation of (99m) Tc-DPD SPECT/CT scan findings and diagnostic blockades of lumbar medial branches in patients with unspecific low back pain in a randomized-controlled trial. *Pain Med*. 2015;16(10):1916–1922.
59 Gornet MG, Peacock J, Claude J, et al. Magnetic resonance spectroscopy (MRS) can identify painful lumbar discs and may facilitate improved clinical outcomes of lumbar surgeries for discogenic pain. *Eur Spine J*. 2019;28(4):674–687.
60 Koh WU, Kim SH, Hwang BY, et al. Value of bone scintigraphy and single photon emission computed tomography (SPECT) in lumbar facet disease and prediction of short-term outcome of ultrasound guided medial branch block with bone SPECT. *Korean J Pain*. 2011;24(2):81–86.
61 Kokkonen SM, Kurunlahti M, Tervonen O, Ilkko E, Vanharanta H. Endplate degeneration observed on magnetic resonance imaging of the lumbar spine: correlation with pain provocation and disc changes observed on computed tomography diskography. *Spine (Phila Pa 1976)*. 2002;27(20):2274–2278.
62 Laslett M, McDonald B, Aprill CN, Tropp H, Oberg B. Clinical predictors of screening lumbar zygapophyseal joint blocks: development of clinical prediction rules. *Spine J*. 2006;6(4):370–379.
63 Laslett M, Oberg B, Aprill CN, McDonald B. Zygapophysial joint blocks in chronic low back pain: a test of Revel's model as a screening test. *BMC Musculoskelet Disord*. 2004;5:43.
64 Manchikanti L, Pampati V, Fellows B, Baha AG. The inability of the clinical picture to characterize pain from facet joints. *Pain Physician*. 2000;3(2):158–166.

65 Manchikanti L, Pampati V, Fellows B, Bakhit CE. Prevalence of lumbar facet joint pain in chronic low back pain. *Pain Physician*. 1999;2(3):59–64.
66 Milette PC, Raymond J, Fontaine S. Comparison of high-resolution computed tomography with discography in the evaluation of lumbar disc herniations. *Spine (Phila Pa 1976)*. 1990;15(6):525–533.
67 Ohnmeiss DD, Vanharanta H, Ekholm J. Relationship of pain drawings to invasive tests assessing intervertebral disc pathology. *Eur Spine J*. 1999;8(2):126–131.
68 Osti OL, Fraser RD. MRI and discography of annular tears and intervertebral disc degeneration. A prospective clinical comparison. *J Bone Joint Surg Br*. 1992;74(3):431–435.
69 Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine (Phila Pa 1976)*. 2001;26(17):1873–1878.
70 Revel M, Poiraudeau S, Auleley GR, et al. Capacity of the clinical picture to characterize low back pain relieved by facet joint anesthesia. Proposed criteria to identify patients with painful facet joints. *Spine (Phila Pa 1976)*. 1998;23(18):1972–1976. discussion 1977.
71 Revel ME, Listrat VM, Chevalier XJ, et al. Facet joint block for low back pain: identifying predictors of a good response. *Arch Phys Med Rehabil*. 1992;73(9):824–828.
72 Simmons JW, Emery SF, McMillin JN, Landa D, Kimmich SJ. Awake discography. A comparison study with magnetic resonance imaging. *Spine (Phila Pa 1976)*. 1991;16(6 Suppl):S216–S221.
73 Vanharanta H, Sachs BL, Spivey M, et al. A comparison of CT/discography, pain response and radiographic disc height. *Spine (Phila Pa 1976)*. 1988;13(3):321–324.
74 Yrjama M, Tervonen O, Kurunlahti M, Vanharanta H. Bony vibration stimulation test combined with magnetic resonance imaging. Can discography be replaced? *Spine (Phila Pa 1976)*. 1997;22(7):808–813.
75 Yrjama M, Tervonen O, Vanharanta H. Ultrasonic imaging of lumbar discs combined with vibration pain provocation compared with discography in the diagnosis of internal anular fissures of the lumbar spine. *Spine (Phila Pa 1976)*. 1996;21(5):571–575.
76 Yrjama M, Vanharanta H. Bony vibration stimulation: a new, non-invasive method for examining intradiscal pain. *Eur Spine J*. 1994;3(4):233–235.
77 Holder LE, Machin JL, Asdourian PL, Links JM, Sexton CC. Planar and high-resolution SPECT bone imaging in the diagnosis of facet syndrome. *J Nucl Med*. 1995;36(1):37–44.
78 Cai G, Laslett LL, Aitken D, et al. Effect of zoledronic acid and Denosumab in patients with low back pain and modic change: a proof-of-principle trial. *J Bone Miner Res*. 2018;33(5):773–782.
79 Peng B, Pang X, Wu Y, Zhao C, Song X. A randomized placebo-controlled trial of intradiscal methylene blue injection for the treatment of chronic discogenic low back pain. *Pain*. 2010;149(1):124–129.
80 Saueressig T, Owen PJ, Diemer F, Zebisch J, Belavy DL. Diagnostic accuracy of clusters of pain provocation tests for detecting sacroiliac joint pain: systematic review with meta-analysis. *J Orthop Sports Phys Ther*. 2021;51(9):422–431.
81 da C Menezes Costa L, Maher CG, Hancock MJ, McAuley JH, Herbert RD, Costa LO. The prognosis of acute and persistent low-back pain: a meta-analysis. *CMAJ*. 2012;184(11):E613–E624.
82 Jacobs JC, Jarvik JG, Chou R, et al. Observational study of the downstream consequences of inappropriate MRI of the lumbar spine. *J Gen Intern Med*. 2020;35(12):3605–3612.

Appendix S1. Search strategy

MEDLINE	EMBASE	CINAHL
1. "Sensitivity and Specificity"/	1. 'Sensitivity and Specificity'/	1. (MH "Sensitivity and Specificity")
2. specificity.mp.	2. Specificity.mp.	2. Specificity
3. False Negative Reactions/ or false negativ*.mp.	3. 'False Negative Results/ OR "false negativ*".mp	3. (MH "False Negative Results") OR "false negativ*"
4. False Positive Reactions/ or false positiv*.mp.	4. 'False Positive Results'/ OR "false positiv*".mp.	4. (MH "False Positive Results") OR "false positiv*"
5. accura*.mp.	5. accura*.mp	5. Valid*
6. "Predictive Value of Tests"/ or predictive value*.mp.	6. 'Predictive Value'/ OR "predictive value*".mp.	6. (MH "Predictive Value of Tests") OR "predictive value*"
7. Reference Values/ or reference value*.mp.	7. 'Reference Value'/ OR "reference value*".mp.	7. (MH "Reference Values") OR "reference value*"
8. (roc adj2 (analys* or area or characteristic* or curve* or estimate* or evaluat*)).mp.	8. (roc adj (analys* OR area OR characteristic* OR curve* OR estimate* OR evaluat*)).mp.	8. "(analys* OR area OR characteristic* OR curve* OR estimate* OR evaluat*))" OR (MH "ROC Curve")
9. ROC Curve/	9. 'Receiver operating characteristic'/	9. (MH "ROC Curve")
10. Likelihood Functions/ or likelihood ratio*.mp.	10. "likelihood ratio*".mp.	10. (MH "Maximum Likelihood ") OR "likelihood ratio*"
11. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10	11. #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10	11. S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10
12. medical history taking.mp. or Medical History Taking/	12. "anamnesis".mp	12. "medical history taking" OR (MH "Patient History Taking")
13. physical exam*.mp. or exp Physical Examination/	13. radiograph*.mp OR 'Radiography'/	13. "physical exam*" OR (MH "Physical Examination+")
14. radiograph*.mp. or exp Radiography/	14. radiograph*.mp OR 'Radiography'/	14. radiograph* OR (MH "Radiography+")
15. x ray*.mp. or X-Rays/	15. "x ray*".mp OR 'X ray/	15. "x ray*" OR (MH "X-Rays")
16. diagnostic procedure*.mp.	16. "diagnostic procedure*".mp	16. "diagnostic procedure*"
17. exp Magnetic Resonance Imaging/ or magnetic resonance imag*.mp.	17. Exp Nuclear Magnetic Resonance Imaging'/ OR "magnetic resonance imag*".mp	17. (MH "Magnetic Resonance Imaging+") OR "magnetic resonance imag*"
18. exp Tomography, X-Ray Computed/ or computed tomograph*.mp.	18. “Computer Assisted Tomography*”.mp	18. (MH "Tomography, X-Ray Computed+") OR "computed tomograph*"
19. nuclear medicine.mp. or Nuclear Medicine/	19. "nuclear medicine".mp OR 'Nuclear Medicine'/	19. "nuclear medicine" OR (MH "Nuclear Medicine")
20. exp Radionuclide Imaging/ or radionuclide imag*.mp.	20. Exp Scintiscanning/	20. (MH "Radionuclide Imaging+") OR "radionuclide imag*"
21. scintigraph*.mp.	21. scintigraph*.mp	21. Radionuclide imag*
22. bone scan*.mp.	22. "Bone scintiscanning ".mp	22. "bone scan*"
23. "Surveys and Questionnaires"/ or questionnaire*.mp.	23. 'Questionnaires'/ OR questionnaire*.mp	23. (MH "Questionnaires") OR questionnaire*
24. clinical history.mp.	24. "clinical history".mp	24. "clinical history"
25. Anesthesia, Spinal/ or spinal injection*.mp.	25. 'Spinal anesthesia'/ OR "spinal injection*".mp	25. (MH "Anesthesia, Spinal") OR "spinal injection*"
26. Diagnostic Tests, Routine/ or diagnostic test*.mp.	26. 'Diagnostic Test,'/ OR "diagnostic test*".mp	26. (MH "Diagnostic Tests, Routine") OR "diagnostic test*"
27. discograph*.mp.	27. discograph*.mp	27. discograph*
28. Thermography/ or thermograph*.mp.	28. 'Thermography'/ OR thermograph*.mp	28. (MH "Thermography") OR thermograph*
29. pain provocation test*.mp.	29. "Provocation test*".mp	29. "pain provocation test*"
30. exp Injections, Intra-Articular/ or intra-articular injection*.mp.	30. 'Intraarticular drug administration/ OR "intraarticular injection*".mp.	30. (MH "Injections, Intraarticular+") OR "intraarticular injection*"
31. exp Nerve Block/ or nerve block*.mp.	31. 'Nerve Block'/ OR "nerve block*".mp	31. (MH "Nerve Block+") OR "nerve block*"
32. diagnostic block*.mp.	32. "diagnostic block*".mp.	32. "diagnostic block*"
33. 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32	33. #12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27 OR #28 OR #29 OR #30 OR #31 OR #32	33. S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27 OR S28 OR S29 OR S30 OR S31 OR S32
34. facet.mp.	34. Facet.mp.	34. Facet
35. Zygapophyseal Joint/ or zygapophy*.mp.	35. 'Zygapophyseal Joint'/ OR zygapophy*.mp.	35. (MH "Zygapophyseal Joint") OR zygapophy*
36. posterior element*.mp.	36. "posterior element*".mp.	36."posterior element*"
37. disc.mp.	37. Disc.mp.	37. Disc
38. discs.mp.	38. Discs.mp.	38. Discs
39. disk*.mp.	39. disk*.mp	39. disk*
40. exp Intervertebral Disc/	40. Exp 'Intervertebral Disk'/	40. (MH "Intervertebral Disk+")
41. Sacroiliac Joint/ or sacroiliac.mp.	41. 'Sacroiliac Joint'/ OR sacroiliac.mp.	41. (MH "Sacroiliac Joint") OR sacroiliac
42. sij.mp.	42. Sij.mp.	42. Sij
43. 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42	43. #34 OR #35 OR #36 OR #37 OR #38 OR #39 OR #40 OR #41 OR #42	43. S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41 OR S42
44. Low Back Pain/ or low* back pain.mp.	44. 'Low Back Pain'/ OR "low* back pain".mp.	44. (MH "Low Back Pain") OR "low* back pain"
45. back pain.mp. or Back Pain/	45. "back pain".mp OR 'Backache'/	45. "back pain" OR (MH "Back Pain")
46. spin* pain.mp.	46. "spinal pain".mp	46. "spin* pain"
47. exp Spinal Diseases/ or spine* disease*.mp.	47. Exp 'Spine Disease'/ OR "spine* disease*".mp	47. (MH "Spinal Diseases+") OR "spine* disease*"
48. lumbar pain.mp.	48. "lumbar pain".mp	48. "lumbar pain"
49. 44 or 45 or 46 or 47 or 48	49. #44 OR #45 OR #46 OR #47 OR #48	49. S44 OR S45 OR S46 OR S47 OR S48
50. 11 and 33 and 43 and 49	50. #11 AND #33 AND #43 AND #49	50. S11 AND S33 AND S43 AND S49

Appendix S2. List of excluded full-text articles and the primary reason for exclusion

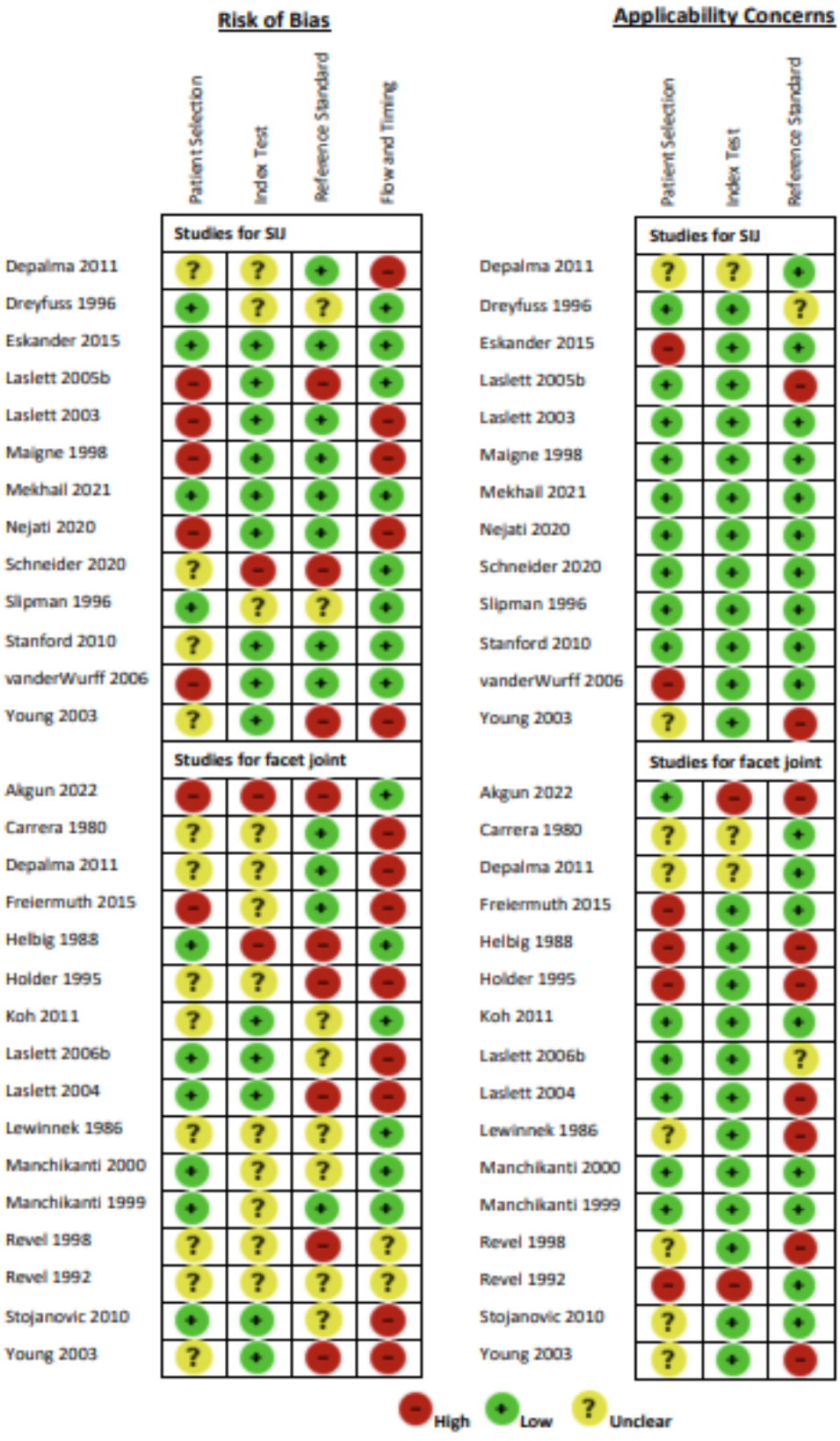
#	Study	Title	First reason for exclusion
1	Acevedo-Gonzalez 2015	Diagnostic test scale SI5: Assessment of sacroiliac joint dysfunction	No 2x2 table or usable data presented
2	Ackerman 2008	Pain Relief With Intraarticular or Medial Branch Nerve Blocks in Patients With Positive Lumbar Facet Joint SPECT Imaging: A 12-Week Outcome Study	No 2x2 table or usable data presented
3	Ahmad 2022	Role of MRI to Rule Out Non Discogenic Causes of Nerve Root Compression	Did not use a reference standard as per protocol
4	Alamin 2011	Provocative lumbar discography versus functional anesthetic discography: a comparison of the results of two different diagnostic techniques in 52 patients with chronic low back pain	Did not assess at least one index test
5	Al-Bedri 2020	The relationship between clinical features and magnetic resonance imaging proved lumbar disc bulging and herniation	Did not use a reference standard as per protocol
6	Al-Bedri 2021	Relationship between disability and radiographic findings of lumbar disk degeneration in sample of Iraqi patients with mechanical low backache	Did not use a reference standard as per protocol
7	Albert 2019	Where do patients with MRI-confirmed single-level radiculopathy experience pain, and what is the clinical interpretability of these pain patterns? A cross-sectional diagnostic accuracy study	Inappropriate population sample
8	Arab 2009	Inter- and intra-examiner reliability of single and composites of selected motion palpation and pain provocation tests for sacroiliac joint	Did not use a reference standard as per protocol
9	Bajada 2016	Psychometric properties including reliability, validity and responsiveness of the Majeed pelvic score in patients with chronic sacroiliac joint pain	Did not use a reference standard as per protocol
10	Bensler 2020	Comparison of treatment outcomes in lumbar disc herniation patients treated with epidural steroid injections: interlaminar versus transforaminal approach	Did not use a reference standard as per protocol
11	Blankenbaker 2006	Axial rotation of the lumbar spinal motion segments correlated with concordant pain on discography: a preliminary study	Did not use a reference standard as per protocol
12	Bogduk 2009	On the rational use of diagnostic blocks for spinal pain	Inappropriate population sample
13	Bogduk 2010	On diagnostic blocks for lumbar zygapophysial joint pain	Inappropriate population sample
14	Bokov 2013	The potential impact of various diagnostic strategies in cases of chronic pain syndromes associated with lumbar spine degeneration	Did not assess at least one index test
15	Borthakur 2011	T1ρ magnetic resonance imaging and discography pressure as novel biomarkers for disc degeneration and low back pain	Did not assess at least one index test
16	Bosscher 2012	Diagnosis of the Vertebral Level from Which Low Back or Leg Pain Originates. A Comparison of Clinical Evaluation, MRI and Epiduroscopy	Did not assess at least one index test
17	Brayda-Bruno 2014	Advances in the diagnosis of degenerated lumbar discs and their possible clinical application	Study is a review
18	Broadhurst 1998	Pain provocation tests for the assessment of sacroiliac joint dysfunction	No 2x2 table or usable data presented
19	Carragee 2006	Low-pressure positive Discography in subjects asymptomatic of significant low back pain illness	Inappropriate population sample
20	Carragee 2006	A gold standard evaluation of the "discogenic pain" diagnosis as determined by provocative discography	Did not use a reference standard as per protocol
21	Chen 2009	Imaging of low back pain: comparative role of high intensity zone in diagnosing the discogenic low back pain with evidence-based radiology	Study is a review
22	Chen 2014	[Magnetic resonance imaging study of disc low back pain]	No 2x2 table or usable data presented
23	Cohen 2007	Clinical predictors of success and failure for lumbar facet radiofrequency denervation	Did not assess at least one index test
24	Cohen 2007	Clinical Predictors of Success and Failure for Lumbar Facet Radiofrequency Denervation	Did not use a reference standard as per protocol
25	Cohen 2013	Facet joint pain--advances in patient selection and treatment	Study is a review
26	Cohen 2013	Establishing an optimal "cutoff" threshold for diagnostic lumbar facet blocks: A prospective correlational study	Did not assess at least one index test
27	Cohen 2014	Can changes in vital signs be used to predict the response to lumbar facet blocks and radiofrequency denervation? A prospective, correlational study	Did not assess at least one index test
28	Cohen 2014	The effect of sedation on the accuracy and treatment outcomes for diagnostic injections: a randomized, controlled, crossover study	Did not assess at least one index test
29	Cohen 2018	Effectiveness of lumbar facet joint blocks and predictive value before radiofrequency denervation the Facet Treatment Study (FACTS), a randomized, controlled clinical trial	Did not assess at least one index test
30	Cohen 2020	Waddell (Nonorganic) Signs and Their Association With Interventional Treatment Outcomes for Low Back Pain	Did not assess at least one index test
31	Cohen 2022	Multicenter study evaluating factors associated with treatment outcome for low back pain injections	Did not use a reference standard as per protocol

32	Conger 2020	The Effectiveness of Radiofrequency Ablation of Medial Branch Nerves for Chronic Lumbar Facet Joint Syndrome in Patients Selected by Guideline-Concordant Dual Comparative Medial Branch Blocks	No 2x2 table or usable data presented
33	Cuellar 2010	Cytokine evaluation in individuals with low back pain using discographic lavage	Did not assess at least one index test
34	DePalma 2012	Multivariable Analyses of the Relationships Between Age, Gender, and Body Mass Index and the Source of Chronic Low Back Pain	Did not assess at least one index test
35	Dephillipo 2019	Sacroiliac pain: Structural causes of pain referring to the si joint region	Did not use a reference standard as per protocol
36	Derby 2012	Comparison of four different analgesic discogram protocols comparing the incidence of reported pain relief following local anesthetic injection into concordantly painful lumbar intervertebral discs	Did not assess at least one index test
37	d'Hemecourt 2019	Sensitivity and specificity of physical examination tests for sacroiliac pain in the adolescent and young athletes	Did not use a reference standard as per protocol
38	Dolan 1996	THE VALUE OF SPECT SCANS IN IDENTIFYING BACK PAIN LIKELY TO BENEFIT FROM FACET JOINT INJECTION	No 2x2 table or usable data presented
39	Eski 2022	Lumbar intervertebral disc degeneration, end-plates and paraspinal muscle changes in children and adolescents with low-back pain	Did not use a reference standard as per protocol
40	Fairbank 1981	Apophyseal injection of local anesthetic as a diagnostic aid in primary low back pain syndromes	Did not use a reference standard as per protocol
41	Ford 2020	Clinical features as predictors of histologically confirmed inflammation in patients with lumbar disc herniation with associated radiculopathy	Did not use a reference standard as per protocol
42	Gupta 2022	Magnetic Resonance Imaging as a Modality for Evaluation of Degenerative Lumbosacral Diseases	Did not use a reference standard as per protocol
43	Gutke 2009	Posterior pelvic pain provocation test is negative in patients with lumbar herniated discs	Did not use a reference standard as per protocol
44	Hao 2006	[The application of lumbar discography in the diagnosis and treatment of the discogenic low back pain]	No 2x2 table or usable data presented
45	Hebelka 2013	HIZ's relation to axial load and low back pain: investigated with axial loaded MRI and pressure controlled discography	Duplicate to Hanna 2013
46	Jackson 1988	Facet joint injection in low back pain: a prospective statistical study	No 2x2 table or usable data presented
47	Jain 2015	Evaluation of Efficacy of Bone Scan With SPECT/CT in the Management of Low Back Pain: A Study Supported by Differential Diagnostic Local Anesthetic Blocks	No 2x2 table or usable data presented
48	Jain 2021	Predictors of discogenic pain in magnetic resonance imaging: a retrospective study of provocative discography performed by posterolateral approach	No 2x2 table or usable data presented
49	Joo 2016	Impact of Type of Needle on Incidence of Intravascular Injection During Diagnostic Lumbar Medial Branch Block	Did not assess at least one index test
50	Juch 2017	Effect of Radiofrequency Denervation on Pain Intensity Among Patients With Chronic Low Back Pain: The Mint Randomized Clinical Trials	No 2x2 table or usable data presented
51	Kasliwal 2016	Fluoroscopy-Guided Sacroiliac Joint Injection: Description of a Modified Technique	Did not assess at least one index test
52	Kluner 2006	Percutaneous discography: comparison of low-dose CT, fluoroscopy and MRI in the diagnosis of lumbar disc disruption	Did not use a reference standard as per protocol
53	Laloo 2018	New bone formation in the intervertebral joint space in spondyloarthritis: An MRI study	Inappropriate population sample
54	Laslett 2006	Pain provocation tests for diagnosis of sacroiliac joint pain	Study is a commentary
55	Laslett 2019	Clinical Diagnosis of Sacroiliac Joint Pain	Study is from a symposium
56	Liliang 2011	Sacroiliac Joint Pain after Lumbar and Lumbosacral Fusion: Findings Using Dual Sacroiliac Joint Blocks	Inappropriate population sample
57	Lilius 1989	The lumbar facet syndrome: a randomised clinical trial	No 2x2 table or usable data presented
58	Liu 2008	The diagnosis of internal disc disruption with CT discography. [Chinese]	No 2x2 table or usable data presented
59	Liu 2018	Painful Schmorl's nodes treated by discography and discoblock	Did not assess at least one index test
60	Manchikanti 2000	The Diagnostic Validity and Therapeutic Value of Lumbar Facet Joint Nerve Blocks with or Without Adjuvant Agents	No 2x2 table or usable data presented
61	Manchikanti 2008	Influence of psychological variables on the diagnosis of facet joint involvement in chronic spinal pain	No 2x2 table or usable data presented
62	Manchikanti 2010	Making sense of the accuracy of diagnostic lumbar facet joint nerve blocks: an assessment of the implications of 50% relief, 80% relief, single block, or controlled diagnostic blocks	No 2x2 table or usable data presented
63	Maus 2014	Diagnostic and therapeutic spinal interventions: Diskography	Study is an editorial

64	Miranda 2021	Sensitivity of Laségue Sign and Slump Test in Hernia and Disc Bulging Diagnoses Compared with Magnetic Resonance Imaging	Did not use a reference standard as per protocol
65	Montes Garcia 2007	[Evocative lumbar discography]	No 2x2 table or usable data presented
66	Ozcan 2021	Is the Distribution Pattern of Modic Changes in Vertebral End-plates Associated With the Severity of Intervertebral Disc Degeneration?: A Cross-sectional Analysis of 527 Caucasians	Did not use a reference standard as per protocol
67	Pampati 2009	Accuracy of diagnostic lumbar facet joint nerve blocks: a 2-year follow-up of 152 patients diagnosed with controlled diagnostic blocks	Did not use a reference standard as per protocol
68	Park 2021	Comparative analysis of the pain provocation test and the HABER test to diagnose nonspecific low-back pain associated with the sacroiliac joint	Did not use a reference standard as per protocol
69	Pesonen 2021	Extending the straight leg raise test for improved clinical evaluation of sciatica: validity and diagnostic performance with reference to the magnetic resonance imaging	Did not use a reference standard as per protocol
70	Pesteie 2015	Real-time ultrasound image classification for spine anesthesia using local directional Hadamard features	Inappropriate population sample
71	Pinto 2022	Provocative Discography: Diagnostic Efficacy and Safety in Symptomatic Degenerative Disk Disease	No 2x2 table or usable data presented
72	Pneumaticos 2006	Low Back Pain: Prediction of Shortterm Outcome of Facet Joint Injection with Bone Scintigraphy	No 2x2 table or usable data presented
73	Putzier 2013	Can discoblock replace discography for identifying painful degenerated discs?	No 2x2 table or usable data presented
74	Richardson 2009	Bilateral L1 and L2 dorsal root ganglion blocks for discogenic low-back pain	Did not assess at least one index test
75	Rocha 2014	Controlled medial branch anesthetic block in the diagnosis of chronic lumbar facet joint pain: the value of a three-month follow-up	Did not assess at least one index test
76	Salman 2022	The Diagnostic Value of Clinical Sacroiliac Joint Pain Provocation Tests for Detection of Sacroiliitis Identified by MRI in Sampled Iraqi Patients with Inflammatory Back Pain	Did not use a reference standard as per protocol
77	Schliessbach 2010	Blockade of the sinuvertebral nerve for the diagnosis of lumbar diskogenic pain: an exploratory study	Did not assess at least one index test
78	Schwarzer 1994	Clinical features of patients with pain stemming from the lumbar zygapophysial joints: Is the lumbar facet syndrome a clinical entity	No 2x2 table or usable data presented
79	Schwarzer 1995	Prevalence and clinical features of lumbar zygapophysial joint pain: a study in an Australian population with chronic low back pain	No 2x2 table or usable data presented
80	Schwarzer 1995a	The ability of computed tomography to identify a painful zygapophysial joint in patients with chronic low back pain	No 2x2 table or usable data presented
81	Shim 2006	Ultrasound-guided lumbar medial-branch block: a clinical study with fluoroscopy control	Did not assess at least one index test
82	Son 2021	Disc height discrepancy between supine and standing positions as a screening metric for discogenic back pain in patients with disc degeneration	No 2x2 table or usable data presented
83	Streitberger 2011	Factors determining the success of radiofrequency denervation in lumbar facet joint pain: a prospective study	Did not assess at least one index test
84	Su 2022	Automatic Grading of Disc Herniation, Central Canal Stenosis and Nerve Roots Compression in Lumbar Magnetic Resonance Image Diagnosis	Did not use a reference standard as per protocol
85	Sudol-Szopinska 2015	Diagnostics of Sacroiliitis According to ASAS Criteria: A Comparative Evaluation of Conventional Radiographs and MRI in Patients with a Clinical Suspicion of Spondyloarthropathy. Preliminary Results	Did not use a reference standard as per protocol
86	Suzuki 2016	Diagnosis and characters of non-specific low back pain in Japan: The Yamaguchi low back pain study	Did not use a reference standard as per protocol
87	Telli 2018	The Validity and Reliability of Provocation Tests in the Diagnosis of Sacroiliac Joint Dysfunction	Did not use a reference standard as per protocol
88	Teraguchi 2020	Lumbar high-intensity zones on MRI: imaging biomarkers for severe, prolonged low back pain and sciatica in a population-based cohort	Did not use a reference standard as per protocol
89	Thompson 2009	Modic changes on MR images as studied with provocative diskography: clinical relevance--a retrospective study of 2457 disks	No 2x2 table or usable data presented
90	Tonosu 2018	Validation study of a diagnostic scoring system for sacroiliac joint-related pain	Did not use a reference standard as per protocol
91	Toren 2020	MRI During Spinal Loading Reveals Intervertebral Disc Behavior Corresponding to Discogram Findings of Annular Fissures and Pain Provocation	No 2x2 table or usable data presented
92	Veizi 2011	Medial branch blocks and facet joint injections as predictors of successful radiofrequency ablation	Did not use a reference standard as per protocol
93	Wang 2008	[Correlation between high intensity zone on MRI and positive pain response on lumbar discography in the diagnosis of discogenic low back pain]	No 2x2 table or usable data presented

94	Wang 2021	Imaging Analysis of the High-Intensity Zone on Lumbar Spine Magnetic Resonance Images: Classification, Features and Correlation with Low Back Pain	Did not use a reference standard as per protocol
95	Weber 2015	Candidate lesion-based criteria for defining a positive sacroiliac joint MRI in two cohorts of patients with axial spondyloarthritis	Did not use a reference standard as per protocol
96	Werner 2013	Distraction test of the posterior superior iliac spine (PSIS) in the diagnosis of sacroiliac joint arthropathy	Inappropriate population sample
97	Yamashita 2019	Accurate diagnosis of low back pain in adult elite athletes	Did not use a reference standard as per protocol
98	Yuan 2006	[Comparison between CT-discography and magnetic resonance imaging in lumbar disc diseases]	No 2x2 table or usable data presented
99	Zhao 2018	Lumbar intervertebral disc degeneration assessed by the nine-point system with X-ray. [Chinese]	Did not use a reference standard as per protocol
100	Zheng 2015	Disc degeneration implies low back pain	Did not use a reference standard as per protocol
101	Zhou 2016	The correlation between radiographic and pathologic grading of lumbar facet joint degeneration	Did not use a reference standard as per protocol
102	Zhu 2016	A method of localization and segmentation of intervertebral discs in spine MRI based on Gabor filter bank	Did not use a reference standard as per protocol
103	Zigler 2018	Progression of Adjacent-level Degeneration After Lumbar Total Disc Replacement: Results of a Post-hoc Analysis of Patients With Available Radiographs From a Prospective Study With 5-year Follow-up	Did not use a reference standard as per protocol
104	Zobel 2012	T1rho magnetic resonance imaging quantification of early lumbar intervertebral disc degeneration in healthy young adults	Did not use a reference standard as per protocol

Appendix S3. Figure 2. Risk of bias and applicability concerns for each domain for each included study for the disc



Appendix S3. Figure 3. Risk of bias and applicability concerns for each domain for each included study for the SIJ and facet

Appendix S4. Methodological quality of included studies

Study	D1. Q1	D1. Q2	D1. Q3	D1. ROB	D1. App	D2. Q1	D2. Q2	D2. ROB	D2. App	D3. Q1	D3. Q2	D3. ROB	D3. App	D4. Q1	D4. Q2	D4. Q3	D4. ROB
Aprill 1992	Yes	Yes	Yes	Low	Low	Yes	Yes	Low	Low	No	Unclear	Low	Low	Unclear	Yes	Yes	Unclear
Akgun 2022	No	Yes	Yes	High	Low	No	Unclear	High	High	No	No	High	High	Yes	Yes	Yes	Low
Bartynski 2023	Yes	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Yes	Yes	Low	Low	No	Yes	Yes	Low
Braithwaite 1998	Yes	Yes	Yes	Low	Low	Yes	Yes	Low	Low	No	No	High	Low	Unclear	Yes	Yes	Unclear
Carragee 2002	Yes	Unclear	Yes	Unclear	Low	Yes	Yes	Low	Low	No	Yes	High	Low	No	Yes	Yes	Low
Carrera 1980	Unclear	Yes	No	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Unclear	Low	Low	Unclear	Yes	No	High
Chelala 2019	No	Yes	Unclear	High	Unclear	No	No	High	Low	No	Unclear	Unclear	Low	No	Yes	Yes	High
Depalma 2011	Yes	Yes	No	Unclear	Unclear	Yes	Unclear	Unclear	Unclear	Yes	No	Low	Low	Unclear	No	No	High
Donelson 1997	Yes	Yes	Yes	Low	High	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Yes	Yes	Yes	Low
Dreyfuss 1996	Yes	Yes	Yes	Low	Low	Unclear	Unclear	Unclear	Low	Yes	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Low
Eskander 2015	Yes	Yes	Yes	Low	High	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Yes	Yes	Yes	Low
Freiermuth 2015	No	Yes	Yes	High	High	Yes	Unclear	Unclear	Low	Yes	Yes	Low	Low	Unclear	Yes	No	High
Gornet 2019	No	Yes	Unclear	High	High	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	High	Unclear	Unclear	No	High
Hanna 2013	Yes	Yes	No	Low	High	Yes	Yes	Low	Low	No	Yes	High	Low	Yes	Yes	No	High
Helbig 1988	Yes	Yes	Unclear	Low	High	No	Unclear	High	Low	No	No	High	High	Unclear	Yes	Yes	Low
Holder 1995	Yes	Unclear	Yes	Unclear	High	Yes	Unclear	Unclear	Low	No	No	High	High	Unclear	No	No	High
Horton 1992	Yes	Yes	No	Low	High	Yes	Unclear	Low	Low	Unclear	Yes	Unclear	Unclear	Unclear	Yes	Yes	Low
Ito 1998	Unclear	Yes	Yes	Unclear	Low	Yes	Yes	Low	Low	Unclear	Yes	High	Unclear	Unclear	Yes	Unclear	Unclear
Kang 2009	Yes	No	No	High	Low	Yes	Yes	Low	Low	Yes	Unclear	Unclear	Low	Unclear	Yes	Yes	Low
Koh 2011	Unclear	Yes	No	Unclear	Low	Yes	Yes	Low	Low	Yes	Unclear	Unclear	Low	Yes	Yes	Yes	Low
Kokkonen 2002	Yes	Yes	Yes	Low	Low	Unclear	Yes	Unclear	Low	No	Unclear	High	High	Unclear	No	Yes	High
Lam 2000	Yes	Yes	Unclear	High	High	Yes	Yes	Low	Low	No	Yes	High	High	Unclear	Yes	No	High
Laslett 2006a	Yes	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Unclear	Yes	Unclear	Unclear	Yes	No	No	High
Laslett 2005b	No	Yes	Yes	High	Low	Yes	Yes	Low	Low	No	Yes	High	High	Yes	Yes	Yes	Low
Laslett 2006b	Yes	Yes	No	Low	Low	Yes	No	Low	Low	Unclear	Yes	Unclear	Unclear	Yes	Yes	No	High
Laslett 2004	Yes	Yes	No	Low	Low	Yes	No	Low	Low	No	Yes	High	High	Unclear	Yes	No	High
Laslett 2005a	Yes	Yes	No	Low	Low	Yes	Unclear	Low	Low	Unclear	Yes	Unclear	Unclear	Yes	Yes	No	High
Laslett 2003	No	Yes	Unclear	High	Low	Yes	Unclear	Low	Low	Yes	Yes	Low	Low	Yes	No	No	High
Lei 2008	Yes	Yes	Yes	Low	High	Yes	Yes	Low	Low	Unclear	No	High	Unclear	Unclear	Yes	Yes	Low
Lewinnek 1986	Unclear	Yes	Unclear	Unclear	Unclear	Unclear	Yes	Unclear	Low	No	Unclear	Unclear	High	Unclear	Yes	Yes	Low
Lim 2005	No	Yes	Unclear	High	Unclear	Yes	Yes	Low	Low	Yes	No	Low	Low	No	Yes	Unclear	Unclear
Maigne 1998	No	No	Yes	High	Low	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Yes	Yes	No	High
Manchikanti 2000	Yes	Yes	No	Low	Low	Unclear	No	Unclear	Low	Yes	Unclear	Unclear	Low	Unclear	Yes	Yes	Low

Manchikanti 1999	Yes	Yes	No	Low	Low	Unclear	No	Unclear	Low	Yes	Yes	Low	Low	Unclear	Yes	Yes	Low
Milette 1990	Unclear	Yes	Yes	Unclear	Low	Yes	No	Low	Low	Unclear	No	Unclear	Unclear	No	Yes	Yes	Unclear
O'Neill 2008	Yes	Yes	Yes	Low	Low	Yes	Yes	Low	Low	No	No	High	High	Unclear	Yes	Yes	Low
Ohnmeiss 1999	Unclear	Yes	Yes	Unclear	Low	Unclear	Yes	Unclear	Low	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Yes	Low
Osti 1992	Yes	Yes	Yes	Low	Low	No	Yes	High	Low	No	No	High	High	Unclear	Yes	Yes	Low
Parker 1996	Yes	Yes	Unclear	Low	High	No	No	High	High	Unclear	No	Unclear	Unclear	Unclear	Yes	Yes	Unclear
Revel 1998	Unclear	Yes	No	Unclear	Unclear	Unclear	Yes	Unclear	Low	No	Unclear	High	High	Unclear	Unclear	Unclear	Unclear
Revel 1992	Unclear	Yes	No	Unclear	High	Unclear	No	Unclear	High	Yes	Unclear	Unclear	Low	Unclear	Yes	No	Unclear
Ricketson 1996	Yes	Yes	Unclear	Unclear	Unclear	Yes	Unclear	Unclear	Unclear	Yes	Yes	Low	Low	Yes	Yes	No	High
Saifuddin 1998	Yes	Yes	No	High	High	Yes	Yes	Low	Low	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Low
Schellhas 1996	Yes	Yes	Unclear	Low	Low	No	Yes	High	Low	Yes	Unclear	Unclear	Low	No	Yes	Yes	High
Schneider 2020	Yes	Yes	Unclear	Unclear	Low	No	No	High	Low	Yes	Unclear	High	Low	Yes	Yes	Yes	Low
Simmons 1991	Yes	Yes	Unclear	Low	Low	Yes	Unclear	Low	Low	Unclear	No	Unclear	Unclear	Unclear	Yes	Yes	Unclear
Slipman 1996	Yes	Yes	No	Low	Low	Unclear	Yes	Unclear	Low	Yes	Unclear	Unclear	Low	Unclear	Yes	Yes	Low
Smith 1998	No	Yes	Unclear	High	High	Unclear	Yes	Unclear	Low	Unclear	Unclear	Unclear	Unclear	No	Yes	Unclear	Unclear
Stanford 2010	Unclear	Yes	Yes	Unclear	Low	Yes	Unclear	Low	Low	Yes	Unclear	Low	Low	Yes	Yes	Yes	Low
Stojanovic 2010	Yes	Yes	Yes	Low	Unclear	Yes	Yes	Low	Low	Yes	Unclear	Unclear	Low	Unclear	No	No	High
vanderWurff 2006	No	Yes	Yes	High	High	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Yes	Yes	Yes	Low
Vanharanta 1998	Unclear	Yes	Unclear	Unclear	Low	Unclear	Yes	Unclear	Low	No	Unclear	High	High	Unclear	Yes	Unclear	Unclear
Vanharanta 1988	Unclear	Yes	Unclear	Unclear	Unclear	Unclear	Yes	Unclear	Low	Unclear	Unclear	Unclear	Unclear	No	Yes	Yes	Low
Weishaupt 2001	Yes	Yes	No	High	High	Yes	Yes	Low	Low	Yes	Unclear	Unclear	Low	No	Yes	Yes	Low
Yoshida 2002	Yes	Yes	No	Low	Unclear	Unclear	Yes	Unclear	Low	Yes	Yes	Low	Low	Unclear	Yes	Yes	Unclear
Young 2003	Unclear	Yes	No	Unclear	Unclear	Yes	Yes	Low	Low	No	Yes	High	High	Unclear	No	No	High
Yrjama 1997	Unclear	Yes	Unclear	Unclear	High	Yes	Yes	Low	Low	Unclear	Yes	Unclear	Unclear	Unclear	Yes	Yes	Low
Yrjama 1996	Unclear	Yes	Unclear	Unclear	Low	No	Yes	High	High	Unclear	Yes	Unclear	Unclear	Yes	Yes	Yes	Low
Yrjama 1994	Unclear	Yes	Unclear	Unclear	Unclear	Yes	No	High	Low	No	Yes	High	High	Yes	Yes	Yes	Low
Nejati 2020	No	Yes	No	High	Low	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Yes	Yes	No	High
Mekhail 2021	Yes	Yes	Unclear	Low	Low	Yes	Yes	Low	Low	Yes	Yes	Low	Low	Yes	Yes	Yes	Low
Waldrop 2021	Unclear	Yes	Yes	Unclear	Low	Yes	Yes	Low	Low	Yes	Unclear	Low	Low	No	Yes	Yes	Low

ROB=Risk of bias; App=Applicability; D=Domain; Q=Signalling question

Appendix S5. Methodological quality of included studies decision guide

Domain 1. Patient selection	Domain 2. Index Test	Domain 3: Reference Standard	Domain 4: Flow and Timing
D1. Signalling question 1: Was a consecutive or random sample of patients enrolled? (Yes/No/Unclear)	**D2. Signalling question 1: Were the index test results interpreted without knowledge of the results of the reference standard? Yes/No/Unclear**	**D3. Signalling question 1: Is the reference standard likely to correctly classify the target condition? Yes/No/Unclear**	**D4. Signalling question 1: Was there an appropriate interval between index test(s) and reference standard? Yes/No/Unclear**
Y= explicitly states that "consecutive series" or "random sample" of participants were enrolled	Y= person conducting and/or interpreting index test index had no knowledge of the results of reference standard (e.g., "the index test is conducted and interpreted before the reference standard", or the index test was conducted blind to the results of the reference standard)	Y= ideal reference standard criteria used from our protocol* (e.g., discogenic pain: discography with pain provocation > 6/10 at disc pressures < 50 psi (if recorded) OR pain behaviours indicative of pain exacerbation), with a single pain-free control disc)	Y= if the "results of the index test and reference standard are collected on the same patients at the same time" (interval between the tests was under 2 weeks for patients with chronic low back pain or under 1 week for acute low back pain)
N= other sampling methods (e.g., convenience sample)	N= person conducting and/or interpreting index test may have had knowledge of the results of reference standard (e.g., the index test was not conducted and interpreted before the reference standard)	N= looser reference standard, for e.g., for discography, researchers used cut-off ≤6/10, no single control disc, higher disc pressures than ideal	N= if the results of the index test and reference standard were not collected on the same patients at the same time (interval between the tests was > 2 weeks for chronic, >1 week for acute)
U= sampling not specified/unclear	U= unclear if reported in the study	U= not clear if reported in the study	U= not clear if reported in the study
D1. Signalling question 2: Was a case-control design avoided? (Yes/No/Unclear)	**D2. Signalling question 2: If a threshold was used, was it pre-specified? Yes/No/Unclear**	**D3. Signalling question 2: Were the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear**	**D4. Signalling question 2: Did all patients receive the same reference standard?**
Y= Used design other than a case-control design	Y= Any threshold for a positive index test was pre-specified e.g., the authors follow a study protocol	Y= person conducting and/or interpreting reference test had no knowledge of results of index test	Y= All patients received the same reference standard
N= Used case control design	N= A threshold was used, and it was not pre-specified (the authors appear to look at lots of different thresholds and pick the best one)	N= person conducting and/or interpreting reference test index had knowledge of results of index test ("Potential for bias is related to the potential influence of previous knowledge on the interpretation of the reference standard")	N= "when only a proportion of the study group receives confirmation of the diagnosis by the reference standard, or if some patients receive a different reference standard."
U= study design not mentioned/unclear	U= unclear if reported in the study	U= not clear if reported in the study	U= not clear if reported in the study
D1. Signalling question 3: Did the study avoid inappropriate exclusions? (Yes/No/Unclear)	**D2. Risk of bias: Could the Conduct or Interpretation of the Index Test Have Introduced Bias?**	**D3. Risk of Bias: Could the Reference Standard, Its Conduct, or its Interpretation Have Introduced Bias?**	**D4. Signalling question 3: Were all patients included in the analysis? Yes/No/Unclear**
Y= The study avoided inappropriate exclusions	Score 'low' if persons conducting/interpreting index text is blinded to the reference test results or conduct of any other index test	Score 'low' if signalling Q1 is rated as 'yes' and blinded interpretation	Y= All (>95%) participants recruited into the study were included in the analysis
N= The study had "inappropriate exclusions" eg., excluding 'difficult to diagnose' patients" or cases where a diagnosis is more obvious than in the usual clinical population (e.g., "excluding patients with "red flags" for the target condition")	Score 'high' if signalling Q1 is scored 'no' or persons conducting/interpreting the index test was not blinded	Score 'high' if signalling Q1 is rated as 'no'	N= "If the number of patients enrolled differs (>5% difference) from the number of patients included in the 2x2 table of results"

U= inclusion/exclusion criteria unclear	Score 'unclear if any of the questions are reported as 'unclear', then there is an unclear risk of bias, and a judgement will be made on whether or not there is enough information to make a decision about the risk of bias	Score 'unclear if any of the questions are reported as 'unclear', then there is an unclear risk of bias, and a judgement will be made on whether or not there is enough information to make a decision about the risk of bias	U= not clear if reported in the study
D1. Risk of bias: Could the Selection of Patients Have Introduced Bias?	**Applicability: Are There Concerns That the Index Test, Its Conduct, or Its Interpretation Differ from the Review Question?**	**Applicability: Are There Concerns That the Target Condition as Defined by the Reference Standard Does Not Match the Question?**	**D4. Risk of Bias: Could the Patient Flow Have Introduced Bias?**
Score 'low' if signalling Q1 and Q2 is rated as 'yes'	Score 'low concerns' if the index text performed is relevant to primary care clinician &/or target condition	Score 'low concern' if reference tests used are relevant to the target condition	Score 'low' if signalling Q2 and Q3 are rated as 'yes'
Score 'high' if signalling Q1 or Q2 is rated as 'no'	Score 'high concerns' if the index test performed is not relevant to primary care clinician &/or target condition	Score 'high concern' if the reference test is not performed to standards in our review protocol	Score 'high' if signalling Q2 or Q3 are rated as 'no'
Score 'unclear if any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement will be made on whether or not there is enough information to make a decision about the risk of bias	Score 'unclear' if insufficient information is provided on how the index test was performed	Score 'unclear' if insufficient information is provided to assess this item	Score 'unclear if any of the questions are reported as 'unclear', then there is an unclear risk of bias, and a judgement will be made on whether or not there is enough information to make a decision about the risk of bias
D1. Applicability: Are There Concerns That the Included Patients and Setting Do Not Match the Review Question? Included patients' criteria: prior testing, presentation, intended use of index test and setting.			
Score 'low concern' if participants are typical of patients from primary care (e.g., persistent non-specific low back pain; failed conservative treatment)			
Score 'high concern' if participants are surgical candidates			
Score 'unclear' if insufficient information is provided to assess this item			

D1=domain 1; D2=domain 2; D3=domain 3; D4=domain 4

Appendix S6. Diagnostic values of all index tests

Author	Tissue source	Index test	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95%CI)	LR- (95%CI)
Aprill et al 1992	Disc	MRI- HIZ vs exact & similar pain reproduction	63.3 (49.9 to 75.4)	96.6 (88.1 to 99.6)	18.36 (4.64 to 72.66)	0.38 (0.27 to 0.53)
Aprill et al 1992	Disc	MRI- HIZ vs exact pain only	81.6 (65.7 to 92.3)	88.8 (79.7 to 94.7)	7.25 (3.85 to 13.66)	0.21 (0.11 to 0.41)
Aprill et al 1992	Disc	MRI- HIZ vs exact & similar pain &morphology	70.6 (56.2 to 82.5)	88.9 (77.4 to 95.8)	6.35 (2.93 to 13.78)	0.33 (0.21 to 0.51)
Akgun et al 2022	Facet joint	Uptake on SPECT	71.4 (55.4 to 84.3)	66.7 (29.9 to 92.5)	2.14 (0.83 to 5.51)	0.43 (0.22 to 0.83)
Bartynski et al 2023	Disc	MRI – Annular fissure	64.0 (49.2 to 77.1)	85.7 (72.8 to 94.1)	4.48 (2.19 to 9.17)	0.42 (0.29 to 0.62)
Braithwaite et al 1998	Disc	MRI- disc degeneration (reduced signal intensity)	86.7 (77.9 to 92.9)	79.0 (66.8 to 88.3)	4.13 (2.53 to 6.75)	0.17 (0.10 to 0.29)
Braithwaite et al 1998	Disc	MRI- Modic changes Grade 1	5.4 (1.8 to 12.1)	99.2 (92.8 to 100.00)	6.72 (0.37 to 120.8)	0.95 (0.90 to 1.01)
Braithwaite et al 1998	Disc	MRI- Modic changes Grade 2	17.2 (10.2 to 26.4)	96.8 (88.8 to 99.6)	5.33 (1.27 to 22.38)	0.86 (0.77 to 0.95)
Braithwaite et al 1998	Disc	MRI- Modic changes Grade 3	3.2 (0.7 to 9.1)	99.2 (92.8 to 100.0)	4.03 (0.21 to 79.12)	0.98 (0.93 to 1.02)
Braithwaite et al 1998	Disc	MRI- Modic changes all grades	23.3 (15.1 to 33.4)	96.8 (88.8 to 99.6)	7.23 (1.76 to 29.74)	0.79 (0.70 to 0.90)
Carragee et al 2002	Disc	MRI- disc degeneration all grades	94.1 (71.3 to 99.9)	43.1 (30.2 to 56.8)	1.65 (1.28 to 2.13)	0.14 (0.02 to 0.94)
Carragee et al 2002	Disc	MRI- disc degeneration grade 2 and 3	82.4 (56.6 to 96.2)	62.1 (48.4 to 74.5)	2.17 (1.46 to 3.23)	0.28 (0.10 to 0.81)
Carragee et al 2002	Disc	MRI-disc degeneration grade 3 only	47.1 (23.0 to 72.2)	89.7 (78.8 to 96.1)	4.55 (1.83 to 11.3)	0.59 (0.37 to 0.93)
Carragee et al 2002	Facet joint	CT – intra-articular facet joint degeneration	88.9 (33.0 to 100.0)	92.3 (47.2 to 100.0)	11.55 (0.79 to 169)	0.12 (0.01 to 1.66)
Chelala et al 2019	Disc	MRI – High intensity zone	24.6 (21.7 to 27.6)	94.6 (93.4 to 95.7)	4.57 (3.61 to 5.79)	0.80 (0.77 to 0.83)
Chelala et al 2019	Disc	MRI – Disc protrusion	22.8 (20.1 to 25.8)	96.7 (95.8 to 97.6)	7.02 (5.23 to 9.43)	0.80 (0.77 to 0.83)
Chelala et al 2019	Disc	MRI – Disc extrusion	4.3 (3.1 to 5.9)	99.8 (99.5 to 100)	22.98 (7.11 to 74.32)	0.96 (0.95 to 0.97)
Chelala et al 2019	Disc	MRI – Disc bulge	13.1 (11 to 15.3)	52.9 (51.1 to 54.7)	0.28 (0.24 to 0.33)	1.65 (1.58 to 1.71)
Chelala et al 2019	Disc	MRI – Disc degeneration	21.9 (19.2 to 24.8)	74.9 (72.7 to 77.0)	0.87 (0.75 to 1.02)	1.04 (1.00 to 1.09)
Chelala et al 2019	Disc	MRI – Disc Listhesis	13.9 (11.6 to 16.4)	96.3 (95.3 to 97.2)	3.76 (2.78 to 5.08)	0.89 (0.87 to 0.92)
Chelala et al 2019	Disc	MRI – Disc herniation	80.3 (75.3 to 84.6)	71.4 (69.4 to 73.3)	2.80 (2.57 to 3.06)	0.28 (0.22 to 0.35)
Depalma et al 2011	Disc	Midline LBP	95.8 (88.1 to 99.1)	74.7 (65.0 to 82.9)	3.79 (2.69 to 5.34)	0.06 (0.02 to 0.17)
Depalma et al 2011	Disc	Para-midline LBP	67.3 (52.9 to 79.7)	9.3 (3.8 to 18.3)	3.50 (1.56 to 7.84)	0.74 (0.61 to 0.91)
Depalma et al 2011	Facet joint	Absence of Midline LBP	15.4 (6.9 to 28.1)	28.0 (20.1 to 37.0)	3.03 (2.22 to 4.13)	0.21 (0.11 to 0.41)
Depalma et al 2011	Facet joint	Para-midline LBP	95 (83.1 to 99.4)	25.3 (16.6 to 35.7)	1.27 (1.10 to 1.47)	0.20 (0.05 to 0.80)
Depalma et al 2011	SIJ	Absence of Midline LBP	12.9 (3.6 to 29.8)	36.0 (28.0 to 44.5)	2.42 (1.87 to 3.14)	0.20 (0.08 to 0.51)
Depalma et al 2011	SIJ	Para-midline LBP	96 (79.6 to 99.9)	22.5 (14.9 to 31.9)	1.24 (1.09 to 1.41)	0.18 (0.03 to 1.25)
Donelson et al 1997	Disc	McKenzie centralisers*	63.9 (46.2 to 79.2)	70.4 (49.8 to 86.2)	2.16 (1.15 to 4.05)	0.51 (0.31 to 0.85)
Donelson et al 1997	Disc	McKenzie centralisers and peripheralisers*	94.4 (81.3 to 99.3)	51.9 (31.9 to 71.3)	1.96 (1.32 to 2.92)	0.11 (0.03 to 0.43)
Dreyfuss et al 1996	SIJ	Better with standing	6.7 (1.4 to 18.3)	97.5 (86.8 to 99.9)	2.67 (0.29 to 24.62)	0.96 (0.87 to 1.05)
Dreyfuss et al 1996	SIJ	Better with walking	13.3 (5.1 to 26.8)	77.5 (61.5 to 89.2)	0.59 (0.23 to 1.52)	1.12 (0.91 to 1.37)
Dreyfuss et al 1996	SIJ	Better with sitting	6.7 (1.4 to 18.3)	85.0 (70.2 to 94.3)	0.44 (0.12 to 1.66)	1.10 (0.94 to 1.28)
Dreyfuss et al 1996	SIJ	Better with lying down	53.3 (37.9 to 68.3)	47.5 (31.5 to 63.9)	1.02 (0.68 to 1.52)	0.98 (0.63 to 1.54)
Dreyfuss et al 1996	SIJ	Worse with cough/sneeze	44.4 (29.6 to 60)	47.5 (31.5 to 63.9)	0.85 (0.55 to 1.31)	1.17 (0.77 to 1.78)
Dreyfuss et al 1996	SIJ	Worse with bowel motion	35.6 (21.9 to 51.2)	62.5 (45.8 to 77.3)	0.95 (0.54 to 1.66)	1.03 (0.75 to 1.43)
Dreyfuss et al 1996	SIJ	Worse with heels/boots	15.6 (6.5 to 29.5)	45.0 (29.3 to 61.5)	0.28 (0.14 to 0.59)	1.88 (1.30 to 2.70)
Dreyfuss et al 1996	SIJ	Worse with job activities	20.0 (9.6 to 34.6)	62.5 (45.8 to 77.3)	0.53 (0.26 to 1.08)	1.28 (0.97 to 1.70)
Dreyfuss et al 1996	SIJ	Sleeps with painful side up	73.3 (58.1 to 85.4)	22.5 (10.8 to 38.5)	0.95 (0.74 to 1.21)	1.19 (0.56 to 2.51)
Dreyfuss et al 1996	SIJ	Pain started with injury	68.9 (53.4 to 81.8)	20.0 (9.1 to 35.6)	0.86 (0.67 to 1.11)	1.56 (0.73 to 3.32)
Dreyfuss et al 1996	SIJ	Better with NSAIDs	45.5 (28.1 to 63.6)	58.8 (40.7 to 75.4)	1.10 (0.64 to 1.91)	0.93 (0.61 to 1.41)

Dreyfuss et al 1996	SIJ	Better with muscle relax	36.0 (18.0 to 57.5)	50.0 (30.6 to 69.4)	0.72 (0.38 to 1.37)	1.28 (0.8 to 2.050)
Dreyfuss et al 1996	SIJ	Better with physical therapy	65.2 (42.7 to 83.6)	80.0 (59.3 to 93.2)	3.26 (1.41 to 7.55)	0.44 (0.24 to 0.79)
Dreyfuss et al 1996	SIJ	Better with home exercises	31.3 (16.1 to 50.0)	83.3 (65.3 to 94.4)	1.88 (0.72 to 4.85)	0.83 (0.62 to 1.10)
Dreyfuss et al 1996	SIJ	Better with heat	61.3 (42.2 to 78.2)	50.0 (32.4 to 67.6)	1.23 (0.79 to 1.90)	0.77 (0.44 to 1.35)
Dreyfuss et al 1996	SIJ	Better with cold	66.7 (44.7 to 84.4)	63.0 (42.4 to 80.6)	1.80 (1.02 to 3.18)	0.53 (0.28 to 1.00)
Dreyfuss et al 1996	SIJ	Better with manipulation	53.6 (33.9 to 72.5)	77.8 (52.4 to 93.6)	2.41 (0.95 to 6.11)	0.60 (0.37 to 0.95)
Dreyfuss et al 1996	SIJ	SIJ pain	80.0 (65.4 to 90.4)	15.0 (5.7 to 29.8)	0.94 (0.77 to 1.15)	1.33 (0.52 to 3.42)
Dreyfuss et al 1996	SIJ	Groin pain	17.8 (8.0 to 32.1)	65.0 (48.3 to 79.4)	0.51 (0.24 to 1.08)	1.27 (0.97 to 1.65)
Dreyfuss et al 1996	SIJ	Buttock pain	82.2 (67.9 to 92)	15.0 (5.7 to 29.8)	0.97 (0.80 to 1.17)	1.19 (0.45 to 3.12)
Dreyfuss et al 1996	SIJ	Sitting position	2.2 (0.1 to 11.8)	82.5 (67.2 to 92.7)	0.13 (0.02 to 0.99)	1.19 (1.02 to 1.38)
Dreyfuss et al 1996	SIJ	PSIS pointing	71.1 (55.7 to 83.6)	47.5 (31.5 to 63.9)	1.35 (0.96 to 1.92)	0.61 (0.35 to 1.07)
Dreyfuss et al 1996	SIJ	Gillet's test	40.0 (25.7 to 55.7)	55.0 (38.5 to 70.7)	0.89 (0.54 to 1.46)	1.09 (0.76 to 1.58)
Dreyfuss et al 1996	SIJ	Thigh thrust	42.2 (27.7 to 57.8)	45.0 (29.3 to 61.5)	0.77 (0.49 to 1.19)	1.28 (0.84 to 1.96)
Dreyfuss et al 1996	SIJ	Patrick's test	57.8 (42.2 to 72.3)	22.5 (10.8 to 38.5)	0.75 (0.55 to 1.01)	1.88 (0.96 to 3.66)
Dreyfuss et al 1996	SIJ	Gaenslen's test	66.7 (51.0 to 80.0)	35.0 (20.6 to 51.7)	1.03 (0.75 to 1.40)	0.95 (0.53 to 1.72)
Dreyfuss et al 1996	SIJ	Midsacral thrust	51.1 (35.8 to 66.3)	40.0 (24.9 to 56.7)	0.85 (0.58 to 1.25)	1.22 (0.75 to 1.98)
Dreyfuss et al 1996	SIJ	Spring	68.9 (53.4 to 81.8)	37.5 (22.7 to 54.2)	1.10 (0.81 to 1.50)	0.83 (0.46 to 1.50)
Dreyfuss et al 1996	SIJ	Sacral sulcus tenderness	93.3 (81.7 to 98.6)	10.0 (2.8 to 23.7)	1.04 (0.91 to 1.18)	0.67 (0.16 to 2.80)
Dreyfuss et al 1996	SIJ	Sacral sulcus tenderness plus PSIS pointing	66.7 (51.0 to 80.0)	50.0 (33.8 to 66.2)	1.33 (0.92 to 1.94)	0.67 (0.40 to 1.12)
Dreyfuss et al 1996	SIJ	Sacral sulcus tenderness plus groin pain	15.6 (6.5 to 29.5)	67.5 (50.9 to 81.4)	0.48 (0.21 to 1.08)	1.25 (0.98 to 1.61)
Dreyfuss et al 1996	SIJ	PSIS pointing plus groin pain	15.6 (6.5 to 29.5)	85.0 (70.2 to 94.3)	1.04 (0.38 to 2.83)	0.99 (0.83 to 1.19)
Dreyfuss et al 1996	SIJ	Sacral sulcus tenderness plus PSIS pointing plus groin pain	13.3 (5.1 to 26.8)	85.0 (70.2 to 94.3)	0.89 (0.31 to 2.54)	1.02 (0.86 to 1.21)
Dreyfuss et al 1996	SIJ	6 positive tests	73.3 (58.1 to 85.4)	20 (9.1 to 35.6)	0.92 (0.73 to 1.16)	1.33 (0.61 to 2.93)
Dreyfuss et al 1996	SIJ	7 positive tests	57.8 (42.2 to 72.3)	45 (29.3 to 61.5)	1.05 (0.72 to 1.53)	0.94 (0.58 to 1.52)
Dreyfuss et al 1996	SIJ	8 positive tests	42.2 (27.7 to 57.8)	50 (33.8 to 66.2)	0.84 (0.53 to 1.34)	1.16 (0.78 to 1.72)
Dreyfuss et al 1996	SIJ	9 positive tests	20 (9.6 to 34.6)	67.5 (50.9 to 81.4)	0.62 (0.30 to 1.28)	1.19 (0.91 to 1.54)
Dreyfuss et al 1996	SIJ	10 positive tests	6.7 (1.4 to 18.3)	87.5 (73.2 to 95.8)	0.53 (0.14 to 2.09)	1.07 (0.93 to 1.23)
Dreyfuss et al 1996	SIJ	11 positive tests	2.2 (0.1 to 11.8)	97.5 (86.8 to 99.9)	0.89 (0.06 to 13.75)	1.00 (0.94 to 1.07)
Eskander et al 2015	SIJ	Distraction test	67 (?)	38 (?)	? (?)	? (?)
Eskander et al 2015	SIJ	FABER test	64 (?)	? (?)	? (?)	? (?)
Eskander et al 2015	SIJ	Gaenslen's test	67 (?)	50 (?)	? (?)	? (?)
Eskander et al 2015	SIJ	Combined physical exam	67 (?)	50 (?)	? (?)	? (?)
Eskander et al 2015	SIJ	Positive fluoroscopic exam	82 (?)	80 (?)	? (?)	? (?)
Freiermuth et al 2015	Facet	SPECT/CT Positive	57.1 (18.4 to 90.1)	77.3 (54.6 to 92.2)	2.51 (0.92 to 6.85)	0.77 (0.55 to 0.92)
Gornet et al 2019	Disc	Magnetic Resonance Spectroscopy (MRS+)	? (?)	? (?)	? (?)	? (?)
Hanna et al 2013	Disc	MRI – High intensity zone	43.7 (31.9 to 56)	72.9 (58.2 to 84.7)	1.61 (0.95 to 2.75)	0.73 (0.58 to 0.85)
Helbig et al 1988	Facet Joint	**Physical examination**				
Helbig et al 1988	Facet Joint	Spasm deformity	? (?)	? (?)	? (?)	? (?)
Helbig et al 1988	Facet Joint	Local paravertebral tenderness	? (?)	? (?)	? (?)	? (?)
Helbig et al 1988	Facet Joint	Diffuse tenderness	? (?)	? (?)	? (?)	? (?)

Helbig et al 1988	Facet Joint	Pain produced by motion	? (?)	? (?)	? (?)	? (?)
Helbig et al 1988	Facet Joint	Minimal or no pain with motion	? (?)	? (?)	? (?)	? (?)
Helbig et al 1988	Facet Joint	SK.LR positive	? (?)	? (?)	? (?)	? (?)
Helbig et al 1988	Facet Joint	**Neurological findings**				
Helbig et al 1988	Facet Joint	Abnormal	72.3 (39.0 to 94.0)	72.3 (39.0 to 94.0)	2.67 (0.95 to 7.47)	0.37 (0.13 to 1.05)
Helbig et al 1988	Facet Joint	**Radiographic findings**				
Helbig et al 1988	Facet Joint	Severe facet changes vs mild + none	60.0 (14.7 to 94.7)	53.3 (26.6 to 78.7)	1.29 (0.52 to 3.15)	0.75 (0.23 to 2.42)
Helbig et al 1988	Facet Joint	3 Positive vs rest	45.5 (16.7 to 76.6)	100.0 (71.5 to 100.0)	11.00 (0.68 to 177.72)	0.57 (0.33 to 0.96)
Helbig et al 1988	Facet Joint	2 Positive	18.2 (2.3 to 51.8)	81.8 (48.2 to 97.7)	1.00 (0.17 to 5.89)	1.00 (0.67 to 1.48)
Helbig et al 1988	Facet Joint	1 Positive	36.4 (10.9 to 69.2)	36.4 (10.9 to 69.2)	0.57 (0.23 to 1.41)	1.75 (0.71 to 4.31)
Holder et al 1995	Facet Joint	Planar dianosis	71.4 (29.0 to 96.3)	76.5 (58.8 to 89.3)	3.04 (1.41 to 6.53)	0.37 (0.11 to 1.22)
Holder et al 1995	Facet Joint	SPECT diagnosis	93.3 (52.4 to 100.0)	70.6 (52.5 to 84.9)	3.17 (1.82 to 5.53)	0.09 (0.01 to 1.39)
Horton et al 1992	Disc	MRI -dark bulged vs pain reproduction	20.0 (5.7 to 43.7)	89.7 (75.8 to 97.1)	1.95 (0.54 to 6.99)	0.89 (0.70 to 1.14)
Horton et al 1992	Disc	MRI- speckled bulged vs pain reproduction	35.0 (15.4 to 59.2)	66.7 (49.8 to 80.9)	1.05 (0.50 to 2.21)	0.98 (0.66 to 1.44)
Horton et al 1992	Disc	MRI- white bulged vs pain reproduction	2.4 (0.0 to 20.6)	87.2 (72.6 to 95.7)	0.19 (0.01 to 3.31)	1.12 (0.97 to 1.29)
Horton et al 1992	Disc	MRI- dark torn vs pain reproduction	15.0 (3.2 to 37.9)	98.7 (88.8 to 100)	11.85 (0.62 to 225.3)	0.86 (0.71 to 1.04)
Horton et al 1992	Disc	MRI- speckled torn vs pain reproduction	10.0 (1.2 to 31.7)	89.7 (75.8 to 97.1)	0.98 (0.20 to 4.98)	1.00 (0.84 to 1.20)
Horton et al 1992	Disc	MRI- dark flat vs pain reproduction	2.4 (0.0 to 20.6)	97.4 (86.5 to 99.9)	0.95 (0.03 to 27.17)	1.00 (0.92 to 1.09)
Horton et al 1992	Disc	MRI- speckled flat vs pain reproduction	10.0 (1.2 to 31.7)	97.4 (86.5 to 99.9)	3.9 (0.38 to 40.45)	0.92 (0.79 to 1.08)
Horton et al 1992	Disc	MRI- white flat vs pain reproduction	10.0 (1.2 to 31.7)	71.8 (55.1 to 85.0)	0.36 (0.09 to 1.45)	1.25 (0.98 to 1.60)
Horton et al 1992	Disc	MRI- dark and speckled vs pain reproduction	90.0 (68.3 to 98.8)	41.0 (25.6 to 57.9)	1.53 (1.13 to 2.06)	0.24 (0.06 to 0.96)
Horton et al 1992	Disc	MRI- dark bulged vs pain reproduction & morphology	21.1 (6.1 to 45.6)	90.0 (76.3 to 97.2)	2.11 (0.59 to 7.53)	0.88 (0.68 to 1.13)
Horton et al 1992	Disc	MRI- speckled bulged vs pain reproduction & morphology	36.8 (16.3 to 61.6)	67.5 (50.9 to 81.4)	1.13 (0.54 to 2.37)	0.94 (0.62 to 1.40)
Horton et al 1992	Disc	MRI- white bulged vs pain reproduction &morphology	2.6 (0.0 to 21.6)	87.5 (73.2 to 95.8)	0.21 (0.01 to 3.57)	1.11 (0.97 to 1.28)
Horton et al 1992	Disc	MRI- dark Torn vs pain reproduction & morphology	15.8 (3.4 to 39.6)	98.8 (89.0 to 100.0)	12.78 (0.67 to 242.8)	0.85 (0.70 to 1.04)
Horton et al 1992	Disc	MRI- speckled torn vs pain reproduction & morphology	10.5 (1.3 to 33.1)	90.0 (76.3 to 97.2)	1.05 (0.21 to 5.25)	0.99 (0.83 to 1.20)
Horton et al 1992	Disc	MRI- dark flat vs pain reproduction & morphology	2.6 (0.0 to 21.6)	97.5 (86.8 to 99.9)	1.03 (0.04 to 29.27)	1.00 (0.92 to 1.09)
Horton et al 1992	Disc	MRI- speckled flat vs pain reproduction & morphology	10.5 (1.3 to 33.1)	97.5 (86.8 to 99.9)	4.21 (0.41 to 43.6)	0.92 (0.78 to 1.08)
Horton et al 1992	Disc	MRI- white flat vs pain reproduction & morphology	5.3 (0.1 to 26.0)	70.0 (53.5 to 83.4)	0.18 (0.03 to 1.25)	1.35 (1.08 to 1.70)
Horton et al 1992	Disc	MRI- dark speckled vs pain reproduction & morphology	94.7 (74.0 to 99.9)	42.5 (27 to 59.1)	1.65 (1.24 to 2.20)	0.12 (0.02 to 0.86)
Ito et al 1998	Disc	MRI- presence of radial tear	87.0 (66.4 to 97.2)	65.4 (53.8 to 75.8)	2.51 (1.78 to 3.54)	0.20 (0.07 to 0.58)
Ito et al 1998	Disc	MRI- massive degeneration	8.7 (1.1 to 28.0)	99.4 (94.2 to 100)	13.65 (0.64 to 292.30)	0.92 (0.81 to 1.04)
Ito et al 1998	Disc	MRI- mod narrowing	56.5 (34.5 to 76.8)	71.8 (60.5 to 81.4)	2.00 (1.21 to 3.32)	0.61 (0.37 to 0.99)
Ito et al 1998	Disc	MRI- severe narrowing	30.4 (13.2 to 52.9)	97.4 (91.0 to 99.7)	11.87 (2.65 to 53.25)	0.71 (0.54 to 0.94)
Ito et al 1998	Disc	MRI- moderate & severe narrowing	87.0 (66.4 to 97.2)	69.2 (57.8 to 79.2)	2.83 (1.96 to 4.09)	0.19 (0.07 to 0.55)
Ito et al 1998	Disc	MRI- moderate signal loss	26.1 (10.2 to 48.4)	57.7 (46.0 to 68.8)	0.62 (0.30 to 1.29)	1.28 (0.94 to 1.74)
Ito et al 1998	Disc	MRI- severe signal loss	69.6 (47.1 to 86.8)	88.5 (79.2 to 94.6)	6.03 (3.08 to 11.79)	0.34 (0.19 to 0.64)
Ito et al 1998	Disc	MRI- moderate & severe signal loss	95.7 (78.1 to 99.9)	46.2 (34.8 to 57.8)	1.78 (1.42 to 2.22)	0.09 (0.01 to 0.65)
Ito et al 1998	Disc	MRI- disruption of posterior outermost annulus	34.8 (16.4 to 57.3)	89.7 (80.8 to 95.5)	3.39 (1.43 to 8.04)	0.73 (0.53 to 0.99)
Ito et al 1998	Disc	MRI- bone marrow intensity change	21.7 (7.5 to 43.7)	94.9 (87.4 to 98.6)	4.24 (1.24 to 14.5)	0.83 (0.66 to 1.03)

Ito et al 1998	Disc	MRI- HIZ	52.2 (30.6 to 73.2)	89.7 (80.8 to 95.5)	5.09 (2.37 to 10.92)	0.53 (0.35 to 0.82)
Kang et al 2009	Disc	MRI – High intensity zone	56.8 (?)	83.6 (?)	? (?)	? (?)
Kang et al 2009	Disc	MRI – Disc protrusion	68.2 (?)	80.6 (?)	? (?)	? (?)
Kang et al 2009	Disc	MRI – Class 2	14.2 (?)	88.6 (?)	? (?)	? (?)
Kang et al 2009	Disc	MRI – Class 4	45.5 (?)	97.8 (?)	? (?)	? (?)
Kang et al 2009	Disc	MRI – Disc degeneration	95.4 (?)	38.8 (?)	? (?)	? (?)
Kang et al 2009	Disc	MRI – Endplate abnormality	13.6 (?)	87.3 (?)	? (?)	? (?)
Koh et al 2011	Facet joint	SPECT 2nd week	96 (79.6 to 99.9)	50 (15.7 to 84.3)	1.92 (0.96 to 3.86)	0.50 (0.16 to 0.84)
Koh et al 2011	Facet joint	SPECT 4th week	97.8 (81 to 100)	50 (18.7 to 81.3)	1.96 (1.05 to 3.65)	0.50 (0.19 to 0.81)
Kokkonen et al 2002	Disc	MRI- endplate degeneration ≥ grade 1	40.5 (24.8 to 57.9)	63.6 (50.9 to 75.1)	1.12 (0.67 to 1.85)	0.93 (0.68 to 1.29)
Kokkonen et al 2002	Disc	MRI- endplate degeneration ≥ grade 2	21.6 (9.8 to 38.2)	78.8 (67 to 87.9)	1.02 (0.47 to 2.2)	1.00 (0.81 to 1.23)
Kokkonen et al 2002	Disc	MRI endplate degeneration ≥ grade 3	2.7 (0.1 to 14.2)	98.5 (91.8 to 100)	1.78 (0.12 to 27.69)	0.99 (0.93 to 1.05)
Lam et al 2000	Disc	MRI- HIZ vs exact/typical pain	78.4 (68.8 to 86.1)	72.4 (59.1 to 83.3)	2.84 (1.85 to 4.37)	0.30 (0.20 to 0.45)
Lam et al 2000	Disc	MRI-HIZ vs exact and similar pain and morphology	80.8 (71.7 to 88)	78.6 (65.6 to 88.4)	3.77 (2.26 to 6.28)	0.24 (0.16 to 0.37)
Laslett et al 2003	SIJ	3 or more positive SIJ pain provocation tests	91.7 (61.5 to 99.8)	79.4 (61.2 to 91.6)	4.44 (2.19 to 9.00)	0.11 (0.02 to 0.69)
Laslett et al 2003	SIJ	3 or more positive SIJ pain provocation tests after excluding suspected discogenic cases using McKenzie examination	91.7 (61.5 to 99.8)	90.9 (70.8 to 98.9)	10.08 (2.66 to 38.21)	0.09 (0.01 to 0.60)
Laslett et al 2004	Facet joint	Revel's criteria vs ≥ 75% pain relief	11.1 (2.4 to 29.2)	91.0 (83.1 to 96.0)	1.24 (0.35 to 4.34)	0.98 (0.84 to 1.13)
Laslett et al 2004	Facet joint	Revel's criteria including recumbancy vs ≥75% pain relief	11.1 (2.4 to 29.2)	91.0 (83.1 to 96.0)	1.24 (0.35 to 4.34)	0.98 (0.84 to 1.13)
Laslett et al 2004	Facet joint	Revel's criteria vs abolition of pain	16.7 (3.6 to 41.4)	93.3 (86.1 to 97.5)	2.50 (0.69 to 9.08)	0.89 (0.72 to 1.11)
Laslett et al 2004	Facet joint	Revel's criteria including recumbancy vs abolition of pain	16.7 (3.6 to 41.4)	92.8 (85.4 to 97.2)	2.32 (0.65 to 8.27)	0.9 (0.73 to 1.11)
Laslett et al 2005a	Disc	Centralisation (all patients with at least partial exam)*	37.7 (25.6 to 51)	90.9 (70.8 to 98.9)	4.15 (1.06 to 16.16)	0.69 (0.54 to 0.87)
Laslett et al 2005a	Disc	Centralisation (full exam)*	40.4 (27 to 54.9)	94.1 (71.3 to 99.9)	6.87 (1.00 to 47.29)	0.63 (0.49 to 0.82)
Laslett et al 2005b	SIJ	Distraction test	47.4 (24.4 to 71.1)	78.6 (59.0 to 91.7)	2.21 (0.94 to 5.19)	0.67 (0.42 to 1.07)
Laslett et al 2005b	SIJ	Thigh thrust	70 (45.7 to 88.1)	64.3 (44.1 to 81.4)	1.96 (1.10 to 3.48)	0.47 (0.23 to 0.96)
Laslett et al 2005b	SIJ	Gaenslen's test right side	47.4 (24.4 to 71.1)	70.4 (49.8 to 86.2)	1.60 (0.76 to 3.39)	0.75 (0.46 to 1.22)
Laslett et al 2005b	SIJ	Gaenslen's test left side	38.9 (17.3 to 64.3)	74.1 (53.7 to 88.9)	1.50 (0.63 to 3.55)	0.83 (0.54 to 1.27)
Laslett et al 2005b	SIJ	Compression	60.0 (36.1 to 80.9)	67.9 (47.6 to 84.1)	1.87 (0.98 to 3.56)	0.59 (0.33 to 1.07)
Laslett et al 2005b	SIJ	Sacral Thrust	55.0 (31.5 to 76.9)	75 (55.1 to 89.3)	2.20 (1.04 to 4.68)	0.60 (0.35 to 1.02)
Laslett et al 2005b	SIJ	1 or more of 6 positive SIJ pain provocation tests	90.0 (68.3 to 98.8)	42.9 (24.5 to 62.8)	1.58 (1.11 to 2.24)	0.23 (0.06 to 0.93)
Laslett et al 2005b	SIJ	2 or more of 6 positive SIJ pain provocation tests	80.0 (56.3 to 94.3)	64.3 (44.1 to 81.4)	2.24 (1.3 to 3.86)	0.31 (0.12 to 0.78)
Laslett et al 2005b	SIJ	3 or more of 6 positive SIJ pain provocation tests	75.0 (50.9 to 91.3)	75.0 (55.1 to 89.3)	3.00 (1.51 to 5.98)	0.33 (0.15 to 0.73)
Laslett et al 2005b	SIJ	4 or more of 6 positive SIJ pain provocation tests	45.0 (23.1 to 68.5)	78.6 (59 to 91.7)	2.10 (0.89 to 4.96)	0.70 (0.45 to 1.09)
Laslett et al 2005b	SIJ	5 or more of 6 positive SIJ pain provocation tests	20.0 (5.7 to 43.7)	85.7 (67.3 to 96)	1.40 (0.40 to 4.94)	0.93 (0.72 to 1.22)
Laslett et al 2005b	SIJ	6 of 6 positive SIJ pain provocation tests	5.0 (0.1 to 24.9)	85.7 (67.3 to 96)	0.35 (0.04 to 2.9)	1.11 (0.92 to 1.33)
Laslett et al 2006a	Disc	Vulnerable in neutral zone*	29.5 (16.8 to 45.2)	81.0 (58.1 to 94.6)	1.55 (0.58 to 4.19)	0.87 (0.66 to 1.15)
Laslett et al 2006a	Disc	Frequent episodes of LBP*	23.3 (11.8 to 38.6)	95.5 (77.2 to 99.9)	5.12 (0.7 to 37.44)	0.80 (0.67 to 0.97)
Laslett et al 2006a	Disc	Persistent pain between episodes*	24.4 (12.4 to 40.3)	97.7 (80.2 to 100)	10.48 (0.64 to 171)	0.77 (0.64 to 0.93)
Laslett et al 2006a	Disc	Complete &partial centralisers*	35.6 (21.9 to 51.2)	88.2 (63.6 to 98.5)	3.02 (0.78 to 11.77)	0.73 (0.55 to 0.96)

Laslett et al 2006a	Disc	Complete centralisers*	22.2 (11.2 to 37.1)	97.1 (76.3 to 100)	7.78 (0.48 to 126.1)	0.80 (0.67 to 0.95)
Laslett et al 2006a	Disc	Directional preference*	48.9 (34.1 to 63.9)	93.3 (68.1 to 99.8)	7.34 (1.08 to 49.86)	0.55 (0.40 to 0.75)
Laslett et al 2006a	Disc	Persistent pain between episodes (PPE)	32.3 (21.2 to 45.1)	92.0 (74.0 to 99.0)	4.04 (1.02 to 15.97)	0.92 (0.74 to 0.99)
Laslett et al 2006a	Disc	Moderate or major loss of extension (ExtLoss)	26.7 (17.1 to 38.1)	87.1 (70.2 to 96.4)	2.07 (0.77 to 5.55)	0.87 (0.70 to 0.96)
Laslett et al 2006a	Disc	Vulnerability in the neutral zone (VABLE)	41.2 (29.4 to 53.8)	83.3 (62.6 to 95.3)	2.47 (0.97 to 6.32)	0.83 (0.63 to 0.95)
Laslett et al 2006b	Facet joint	Painful extension/rotation vs 75% pain relief	83.3 (62.6 to 95.3)	22.0 (12.3 to 34.7)	1.07 (0.85 to 1.34)	0.76 (0.27 to 2.09)
Laslett et al 2006b	Facet joint	Painful extension/rotation vs 80% pain relief	85.7 (63.7 to 97)	22.6 (12.9 to 35)	1.11 (0.89 to 1.38)	0.63 (0.20 to 1.99)
Laslett et al 2006b	Facet joint	Painful extension/rotation vs 85% pain relief	88.9 (65.3 to 98.6)	23.1 (13.5 to 35.2)	1.16 (0.94 to 1.43)	0.48 (0.12 to 1.91)
Laslett et al 2006b	Facet joint	Painful extension/rotation vs 90% pain relief	92.3 (64 to 99.8)	22.9 (13.7 to 34.4)	1.2 (0.98 to 1.47)	0.34 (0.05 to 2.32)
Laslett et al 2006b	Facet joint	Painful extension/rotation vs 95% pain relief	95.2 (63.3 to 100)	23.3 (14.2 to 34.6)	1.24 (1.03 to 1.49)	0.20 (0.01 to 3.16)
Laslett et al 2006b	Facet joint	Painful extension/rotation vs 100% pain relief	95.2 (63.3 to 100)	23.3 (14.2 to 34.6)	1.24 (1.03 to 1.49)	0.20 (0.01 to 3.16)
Laslett et al 2006b	Facet joint	Painful ipsilateral extension/rotation vs 95% pain relief	87.5 (47.3 to 99.7)	32.7 (20.7 to 46.7)	1.3 (0.94 to 1.79)	0.38 (0.06 to 2.48)
Laslett et al 2006b	Facet joint	Painful contralateral extension/rotation vs 95% pain relief	50 (15.7 to 84.3)	38.2 (25.4 to 52.3)	0.81 (0.39 to 1.67)	1.31 (0.61 to 2.83)
Laslett et al 2006b	Facet joint	Age >50 vs 95% pain relief	54.5 (23.4 to 83.3)	77.3 (66.2 to 86.2)	2.41 (1.22 to 4.76)	0.59 (0.30 to 1.14)
Laslett et al 2006b	Facet joint	Best activity is sitting vs 95% pain relief	18.2 (2.3 to 51.8)	89 (79.5 to 95.1)	1.66 (0.40 to 6.82)	0.92 (0.69 to 1.23)
Laslett et al 2006b	Facet joint	Best activity is walking vs 95% pain relief	36.4 (10.9 to 69.2)	89.2 (79.8 to 95.2)	3.36 (1.21 to 9.32)	0.71 (0.45 to 1.12)
Laslett et al 2006b	Facet joint	MSPQ 13 or more vs 95% pain relief	45.5 (16.7 to 76.6)	65.8 (53.7 to 76.5)	1.33 (0.65 to 2.73)	0.83 (0.47 to 1.46)
Laslett et al 2006b	Facet joint	Absence of centralisation vs 95% pain relief	95.2 (63.3 to 100)	15.8 (7.5 to 27.9)	1.13 (0.95 to 1.35)	0.30 (0.02 to 4.82)
Laslett et al 2006b	Facet joint	Absence of directional preference vs 95% pain relief	88.9 (51.8 to 99.7)	41.7 (29.1 to 55.1)	1.52 (1.11 to 2.09)	0.27 (0.04 to 1.73)
Laslett et al 2006b	Facet joint	Onset is paraspinal vs 95% pain relief	70 (34.8 to 93.3)	74.3 (62.8 to 83.8)	2.73 (1.56 to 4.78)	0.40 (0.16 to 1.05)
Laslett et al 2006b	Facet joint	CPR1 vs 95% pain relief	90 (55.5 to 99.7)	85.1 (74.3 to 92.6)	6.03 (3.28 to 11.07)	0.12 (0.02 to 0.76)
Laslett et al 2006b	Facet joint	CPR2 vs 95% pain relief	95.2 (63.3 to 100)	37 (26 to 49.1)	1.51 (1.21 to 1.89)	0.13 (0.01 to 1.96)
Laslett et al 2006b	Facet joint	CPR3 vs 95% pain relief	95.2 (63.3 to 100)	26.7 (17.1 to 38.1)	1.30 (1.07 to 1.57)	0.18 (0.01 to 2.74)
Laslett et al 2006b	Facet joint	CPR4 vs 95% pain relief	95.2 (63.3 to 100)	51.4 (39.4 to 63.1)	1.96 (1.49 to 2.57)	0.09 (0.01 to 1.40)
Laslett et al 2006b	Facet joint	CPR5 vs 95% pain relief	81.8 (48.2 to 97.7)	91.8 (83 to 96.9)	9.96 (4.40 to 22.5)	0.20 (0.06 to 0.70)
Laslett et al 2006b	Facet	Age 50 or more	61.5 (31.6 to 86.1)	76.6 (67.5 to 84.3)	2.63 (1.52 to 4.57)	0.77 (0.68 to 0.84)
Laslett et al 2006b	Facet	Pain is best when walking	30.8 (9.1 to 61.4)	91.5 (84.5 to 96)	3.62 (1.30 to 10.12)	0.92 (0.85 to 0.96)
Laslett et al 2006b	Facet	Pain is best when sitting	33.3 (9.9 to 65.1)	89.5 (82 to 94.7)	3.18 (1.20 to 8.45)	0.90 (0.82 to 0.95)
Laslett et al 2006b	Facet	Onset of pain is para-spinal	75 (42.8 to 94.5)	72.1 (62.5 to 80.5)	2.69 (1.72 to 4.22)	0.72 (0.63 to 0.81)
Laslett et al 2006b	Facet	MSPQ>13 (somatization)	46.2 (19.2 to 74.9)	69.5 (59.8 to 78.1)	1.51 (0.79 to 2.91)	0.70 (0.60 to 0.78)
Laslett et al 2006b	Facet	Extension/Rotation test	96 (68.3 to 100)	22.3 (14.7 to 31.6)	1.24 (1.06 to 1.44)	0.22 (0.15 to 0.32)
Laslett et al 2006b	Facet	Absence of centralization during repeated movement testing	95.7 (66 to 100)	17.3 (9.8 to 27.3)	1.16 (0.99 to 1.36)	0.17 (0.10 to 0.27)
Lei et al 2008	Disc	MRI – High intensity zone	24.5 (13.8 to 38.3)	86.7 (75.4 to 94.1)	1.84 (0.83 to 4.09)	0.87 (0.75 to 0.94)
Lei et al 2008	Disc	MRI – Endplate changes	32.1 (19.9 to 46.3)	98.3 (91.1 to 100)	19.24 (2.65 to 139.76)	0.98 (0.91 to 1.00)
Lei et al 2008	Disc	MRI pain predictors	94.3 (84.3 to 98.8)	76.7 (64 to 86.6)	4.04 (2.54 to 6.43)	0.77 (0.64 to 0.87)
Lewinnek et al 1986	Facet Joint	Facet joiint DJD on X-ray	81.3 (54.4 to 96.0)	33.3 (0.0 to 98.3)	1.22 (0.38 to 3.87)	0.56 (0.05 to 6.73)
Lewinnek et al 1986	Facet Joint	Wakes from sleep	58.3 (27.7 to 84.8)	94.1 (56.7 to 100.0)	9.92 (0.65 to 152.24)	0.44 (0.22 to 0.88)
Lewinnek et al 1986	Facet Joint	Acute onset	95.2 (63.3 to 100.0)	50.0 (18.7 to 81.3)	1.91 (1.01 to 3.59)	0.10 (0.01 to 1.53)

Lewinnek et al 1986	Facet Joint	Pain below knee	88.9 (51.8 to 99.7)	63.6 (30.8 to 89.1)	2.44 (1.08 to 5.52)	0.18 (0.03 to 1.17)
Lewinnek et al 1986	Facet Joint	Sitting increases pain	63.6 (30.8 to 89.1)	83.3 (35.9 to 99.6)	3.82 (0.60 to 24.14)	0.44 (0.19 to 1.03)
Lewinnek et al 1986	Facet Joint	Time-dependent positional distress	78.6 (49.2 to 95.3)	66.7 (22.3 to 95.7)	2.36 (0.74 to 7.55)	0.32 (0.10 to 1.02)
Lewinnek et al 1986	Facet Joint	NSAID helped	50.0 (11.8 to 88.2)	91.7 (61.5 to 99.8)	6.00 (0.78 to 46.14)	0.55 (0.24 to 1.24)
Lewinnek et al 1986	Facet Joint	Pain on extension	73.3 (44.9 to 92.2)	80.0 (28.4 to 99.5)	3.67 (0.62 to 21.73)	0.33 (0.13 to 0.86)
Lewinnek et al 1986	Facet Joint	SLR	66.7 (22.3 to 95.7)	76.9 (46.2 to 95.0)	2.89 (0.92 to 9.06)	0.43 (0.13 to 1.40)
Lim et al 2005	Disc	MRI- disc degeneration Grades IV & V	88.2 (72.5 to 96.7)	52.4 (39.4 to 65.1)	1.85 (1.39 to 2.47)	0.23 (0.09 to 0.58)
Lim et al 2005	Disc	MRI- HIZ	55.9 (37.9 to 72.8)	69.8 (57.0 to 80.8)	1.85 (1.15 to 3.00)	0.63 (0.42 to 0.95)
Lim et al 2005	Disc	MRI- decreases disc height	29.4 (15.1 to 47.5)	82.5 (70.9 to 90.9)	1.68 (0.80 to 3.56)	0.86 (0.67 to 1.09)
Lim et al 2005	Disc	MRI- abnormal end plate	8.8 (1.9 to 23.7)	82.5 (70.9 to 90.9)	0.51 (0.15 to 1.69)	1.11 (0.95 to 1.29)
Lim et al 2005	Disc	MRI- facet osteoarthritis	55.9 (37.9 to 72.8)	39.7 (27.6 to 52.8)	0.93 (0.65 to 1.33)	1.11 (0.68 to 1.81)
Maigne et al 1998	SIJ	Radionuclide imaging (bone scan)	46.2 (19.2 to 74.9)	94.7 (74 to 99.9)	8.77 (1.19 to 64.53)	0.57 (0.34 to 0.95)
Manchikanti et al 1999	Facet joint	Gender- Male	38.9 (25.9 to 53.1)	56.1 (43.3 to 68.3)	0.89 (0.58 to 1.36)	1.09 (0.81 to 1.47)
Manchikanti et al 1999	Facet joint	Age > 65	18.5 (9.3 to 31.4)	66.7 (54 to 77.8)	0.56 (0.29 to 1.07)	1.22 (0.99 to 1.51)
Manchikanti et al 1999	Facet joint	Traumatic	53.7 (39.6 to 67.4)	47 (34.6 to 59.7)	1.01 (0.72 to 1.42)	0.99 (0.67 to 1.45)
Manchikanti et al 1999	Facet joint	Previous surgery	16.7 (7.9 to 29.3)	66.7 (54 to 77.8)	0.5 (0.25 to 0.99)	1.25 (1.02 to 1.54)
Manchikanti et al 2000	Facet joint	Age ≥65	21.4 (13.2 to 31.7)	85.3 (77.6 to 91.2)	1.46 (0.8 to 2.67)	0.92 (0.81 to 1.05)
Manchikanti et al 2000	Facet joint	Pain well relieved by supine	94 (86.7 to 98)	16.4 (10.2 to 24.4)	1.13 (1.02 to 1.24)	0.36 (0.14 to 0.93)
Manchikanti et al 2000	Facet joint	Pain not exacerbated by coughing	90.5 (82.1 to 95.8)	12.9 (7.4 to 20.4)	1.04 (0.94 to 1.15)	0.74 (0.33 to 1.66)
Manchikanti et al 2000	Facet joint	Pain not exacerbated by forward flexion	15.5 (8.5 to 25)	81.9 (73.7 to 88.4)	0.86 (0.45 to 1.61)	1.03 (0.91 to 1.17)
Manchikanti et al 2000	Facet joint	Pain not exacerbated by deflexion	54.5 (43.6 to 65.2)	48.4 (39.2 to 57.6)	1.06 (0.82 to 1.37)	0.94 (0.70 to 1.26)
Manchikanti et al 2000	Facet joint	Pain not exacerbated by hyperextension	9.5 (4.2 to 17.9)	86.2 (78.6 to 91.9)	0.69 (0.31 to 1.54)	1.05 (0.95 to 1.16)
Manchikanti et al 2000	Facet joint	Pain not exacerbated by extension/rotation	67.9 (56.8 to 77.6)	30.2 (22 to 39.4)	0.97 (0.80 to 1.18)	1.07 (0.70 to 1.62)
Manchikanti et al 2000	Facet joint	Traumatic onset	47.6 (36.6 to 58.8)	50 (40.6 to 59.4)	0.95 (0.71 to 1.27)	1.05 (0.80 to 1.38)
Manchikanti et al 2000	Facet joint	Revel's criteria	12.9 (6.6 to 22)	84.3 (76.4 to 90.5)	0.83 (0.41 to 1.66)	1.03 (0.92 to 1.16)
Manchikanti et al 2000	Facet joint	Pain with flexion	84.5 (75.0 to 91.5)	18.1 (11.6 to 26.3)	1.03 (0.91 to 1.17)	0.86 (0.45 to 1.61)
Manchikanti et al 2000	Facet joint	Pain with deflexion	57.1 (45.9 to 67.9)	45.7 (36.4 to 55.2)	1.05 (0.82 to 1.35)	0.94 (0.68 to 1.29)
Manchikanti et al 2000	Facet joint	Pain with extension	88.0 (79 to 94.1)	13.7 (8.0 to 21.3)	1.02 (0.92 to 1.13)	0.88 (0.42 to 1.84)
Manchikanti et al 2000	Facet joint	Pain with lateral rotation	67.9 (56.8 to 77.6)	30.2 (22.0 to 39.4)	0.97 (0.80 to 1.18)	1.07 (0.7.0 to 1.62)
Manchikanti et al 2000	Facet joint	Pain with sitting and bending	48.8 (37.7 to 60)	56.0 (46.5 to 65.2)	1.11 (0.82 to 1.5)	0.91 (0.70 to 1.19)
Manchikanti et al 2000	Facet joint	Normal gait	97.6 (91.7 to 99.7)	10.3 (5.5 to 17.4)	1.09 (1.02 to 1.17)	0.23 (0.05 to 1.00)
Manchikanti et al 2000	Facet joint	Muscle spasm	47.6 (36.6 to 58.8)	45.7 (36.4 to 55.2)	0.88 (0.66 to 1.16)	1.15 (0.86 to 1.52)
Manchikanti et al 2000	Facet joint	Paravertebral tenderness	92.9 (85.3 to 97.4)	13.0 (7.5 to 20.6)	1.07 (0.98 to 1.17)	0.54 (0.22 to 1.34)
Manchikanti et al 2000	Facet joint	Pain increased by cough/valsalva	14.3 (7.6 to 23.6)	86.2 (78.6 to 91.9)	1.04 (0.52 to 2.07)	0.99 (0.89 to 1.11)
Manchikanti et al 2000	Facet joint	Back pain with SLR	49.4 (38.2 to 60.6)	36.8 (28.0 to 46.2)	0.78 (0.60 to 1.01)	1.38 (1.00 to 1.89)
Manchikanti et al 2000	Facet joint	Negative neurological exam	90.6 (82.3 to 95.8)	25.2 (17.6 to 34.2)	1.21 (1.07 to 1.37)	0.37 (0.18 to 0.78)
Manchikanti et al 2000	Facet joint	Facet joint arthritis	37.3 (27 to 48.7)	65.0 (55.6 to 73.5)	1.07 (0.74 to 1.55)	0.96 (0.78 to 1.19)
Manchikanti et al 2000	Facet joint	Osteoporosis	44 (33.2 to 55.3)	74.1 (65.2 to 81.8)	1.70 (1.15 to 2.52)	0.76 (0.61 to 0.94)
Manchikanti et al 2000	Facet joint	> 3 inappropriate symptoms	23.8 (15.2 to 34.3)	68.1 (58.8 to 76.4)	0.75 (0.47 to 1.19)	1.12 (0.94 to 1.33)
Manchikanti et al 2000	Facet joint	> 3 inappropriate signs	53.6 (42.4 to 64.5)	58.6 (49.1 to 67.7)	1.30 (0.97 to 1.74)	0.79 (0.60 to 1.04)
Manchikanti et al 2000	Facet joint	Old age	25 (16.2 to 35.6)	84.5 (76.6 to 90.5)	1.61 (0.92 to 2.83)	0.89 (0.77 to 1.03)

Manchikanti et al 2000	Facet joint	Prior history of LBP	69.9 (58.8 to 79.5)	24.8 (17.3 to 33.6)	0.93 (0.78 to 1.11)	1.22 (0.77 to 1.92)
Manchikanti et al 2000	Facet joint	Post surgery	20 (12.1 to 30.1)	63.5 (54 to 72.3)	0.55 (0.34 to 0.89)	1.26 (1.06 to 1.50)
Manchikanti et al 2000	Facet joint	Onset occupational	14 (7.4 to 23.1)	71.7 (62.7 to 79.5)	0.49 (0.27 to 0.90)	1.2 (1.04 to 1.38)
Manchikanti et al 2000	Facet joint	Positive work status	34 (21.5 to 48.3)	61.8 (50 to 72.8)	0.89 (0.56 to 1.43)	1.07 (0.82 to 1.39)
Manchikanti et al 2000	Facet joint	Duration < 1 year	21.4 (13.2 to 31.7)	79.3 (70.8 to 86.3)	1.04 (0.6 to 1.78)	0.99 (0.86 to 1.15)
Manchikanti et al 2000	Facet joint	Body mass index (normal)	40 (29.5 to 51.2)	68.7 (59.4 to 77)	1.28 (0.88 to 1.86)	0.87 (0.71 to 1.08)
Manchikanti et al 2000	Facet joint	Low back, hip buttocks pain	97.6 (91.6 to 99.7)	0.4 (0.0 to 3.9)	0.98 (0.95 to 1.02)	5.66 (0.26 to 123.9)
Manchikanti et al 2000	Facet joint	Bilateral pain	75.9 (65.3 to 84.6)	35 (26.5 to 44.4)	1.17 (0.98 to 1.40)	0.69 (0.44 to 1.08)
Manchikanti et al 2000	Facet joint	Back pain with groin or thigh	34.5 (24.5 to 45.7)	68.4 (59.1 to 76.8)	1.09 (0.73 to 1.63)	0.96 (0.78 to 1.17)
Manchikanti et al 2000	Facet joint	Pseudoradicular pain	51.2 (40 to 62.3)	43 (33.7 to 52.6)	0.90 (0.69 to 1.17)	1.14 (0.84 to 1.54)
Manchikanti et al 2000	Facet joint	leg pain	58.8 (47.6 to 69.4)	33.9 (25.3 to 43.3)	0.89 (0.71 to 1.11)	1.21 (0.85 to 1.74)
Manchikanti et al 2000	Facet joint	subjective pain ≥8/10	38.1 (27.7 to 49.3)	65.5 (56.1 to 74.1)	1.11 (0.76 to 1.6)	0.95 (0.76 to 1.17)
Manchikanti et al 2000	Facet joint	cramping pain above knees	43.2 (32.2 to 54.7)	56.3 (46.9 to 65.4)	0.99 (0.72 to 1.37)	1.01 (0.79 to 1.29)
Manchikanti et al 2000	Facet joint	Parasthesia	73.8 (63.1 to 82.8)	16.4 (10.2 to 24.4)	0.88 (0.76 to 1.03)	1.60 (0.93 to 2.76)
Manchikanti et al 2000	Facet joint	Low back stiffness	57.8 (46.5 to 68.6)	38.5 (29.6 to 47.9)	0.94 (0.74 to 1.19)	1.10 (0.78 to 1.54)
Mekhail et al 2021	SIJ	Mekhail test	80.4 (73.3 to 86.3)	26.8 (14.2 to 42.9)	1.10 (0.90 to 1.34)	0.73 (0.40 to 1.33)
Mekhail et al 2021	SIJ	Patrick test	81.6 (74.7 to 87.3)	43.9 (28.5 to 60.3)	1.46 (1.10 to 1.93)	0.42 (0.26 to 0.67)
Mekhail et al 2021	SIJ	Thigh thrust test	53.2 (45.1 to 61.1)	48.8 (32.9 to 64.9)	1.04 (0.74 to 1.45)	0.96 (0.67 to 1.37)
Mekhail et al 2021	SIJ	Mekhail/Patrick test	94.3 (89.5 to 97.4)	17.1 (7.2 to 32.1)	1.14 (0.99 to 1.31)	0.34 (0.13 to 0.84)
Mekhail et al 2021	SIJ	Mekhail/thigh thrust test	89.9 (84.1 to 94.1)	14.6 (5.6 to 29.2)	1.05 (0.92 to 1.21)	0.69 (0.29 to 1.66)
Mekhail et al 2021	SIJ	Patrick/thigh thrust test	89.2 (83.3 to 93.6)	29.3 (16.1 to 45.5)	1.26 (1.03 to 1.55)	0.37 (0.19 to 0.71)
Milette et al 1990	Disc	CT- hernia	46.5 (36.5 to 56.7)	79.8 (71.5 to 86.6)	2.31 (1.53 to 3.49)	0.67 (0.55 to 0.82)
Milette et al 1990	Disc	CT degeneration	4.0 (1.1 to 9.8)	97.5 (92.8 to 99.5)	1.57 (0.36 to 6.86)	0.99 (0.94 to 1.04)
Milette et al 1990	Disc	CT hernia or degeneration	50.5 (40.4 to 60.6)	77.3 (68.7 to 84.5)	2.23 (1.52 to 3.27)	0.64 (0.51 to 0.80)
Nejati et al 2020	SIJ	FABER Test	71.8 (55.1 to 85.0)	66.7 (29.9 to 92.5)	2.15 (0.84 to 5.54)	0.42 (0.21 to 0.84)
Nejati et al 2020	SIJ	Thigh thurst test	74.4 (57.9 to 87.0)	44.4 (13.7 to 78.8)	1.34 (0.73 to 2.47)	0.58 (0.23 to 1.43)
Nejati et al 2020	SIJ	Gaenslen test	38.5 (23.4 to 55.4)	33.3 (7.5 to 70.1)	0.58 (0.31 to 1.06)	1.85 (0.71 to 4.81)
Nejati et al 2020	SIJ	Yeoman test	64.1 (47.2 to 78.8)	33.3 (7.5 to 70.1)	0.96 (0.57 to 1.61)	1.08 (0.39 to 2.97)
Nejati et al 2020	SIJ	Gillet test	98.7 (88.8 to 100.00)	5.3 (0.0 to 39.7)	1.04 (0.89 to 1.22)	0.24 (0.01 to 11.36)
Nejati et al 2020	SIJ	Forward flexion test	98.7 (88.8 to 100.00)	5.3 (0.0 to 39.7)	1.04 (0.89 to 1.22)	0.24 (0.01 to 11.36)
Nejati et al 2020	SIJ	3 or more of 6 positive SIJ pain provocation tests	98.7 (88.8 to 100.00)	5.9 (0.0 to 43.3)	1.05 (0.88 to 1.25)	0.23 (0.01 to 10.64)
Nejati et al 2020	SIJ	4 or more of 6 positive SIJ pain provocation tests	92.3 (79.1 to 98.4)	22.2 (2.8 to 60.0)	1.19 (0.83 to 1.70)	0.35 (0.07 to 1.78)
Nejati et al 2020	SIJ	5 or more of 6 positive SIJ pain provocation tests	59.0 (42.1 to 74.4)	55.6 (21.2 to 86.3)	1.33 (0.61 to 2.88)	0.74 (0.37 to 1.48)
Nejati et al 2020	SIJ	Six positive tests	25.6 (13.0 to 42.1)	88.9 (51.8 to 99.7)	2.31 (0.34 to 15.80)	0.84 (0.62 to 1.12)
Nejati et al 2020	SIJ	FABER and Thigh Thrust	71.7 (?)	66.0 (?)	? (?)	? (?)
Nejati et al 2020	SIJ	FABER and Gaenslen	48.7 (?)	77.8 (?)	? (?)	? (?)
Nejati et al 2020	SIJ	FABER and Yeoman	43.6 (?)	66.7 (?)	? (?)	? (?)
Nejati et al 2020	SIJ	Thigh Thrust and Gaenslen	48.7 (?)	66.7 (?)	? (?)	? (?)
Nejati et al 2020	SIJ	Thigh Thrust and Yeoman	43.6 (?)	55.6 (?)	? (?)	? (?)
Nejati et al 2020	SIJ	Gaenslen and Yeoman	35.9 (?)	55.6 (?)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Moderate loss of Nuclear signal	89.0 (85 to 93)	67.0 (58.2 to 75.8)	? (?)	? (?)

O'Neill et al 2008	Disc	MRI – Severe loss of Nuclear signal	24.0 (17.3 to 30.7)	96.0 (79.0 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Moderate loss of disc height	73.0 (67 to 79)	81.0 (69.2 to 92.8)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Severe loss of disc height	18.0 (12.2 to 23.8)	98.0 (80.3 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – High intensity zone (Grade 1)	44.0 (37.5 to 50.5)	89.0 (74.7 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – High intensity zone (Grade 2)	26.0 (20.1 to 31.9)	95.0 (78.4 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – High intensity zone (Grade 3)	15.0 (9.7 to 20.3)	98.0 (80.3 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Bone marrow intensity change (type 1)	14.0 (9.4 to 18.6)	98.0 (80.4 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Bone marrow intensity change (type 2)	7.0 (3.4 to 10.6)	99.0 (81.1 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Disc bulge	74.0 (68.0 to 80.0)	84.0 (71.2 to 96.8)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Disc protrusion	35.0 (28.5 to 41.5)	91.0 (75.8 to 100.0)	? (?)	? (?)
O'Neill et al 2008	Disc	MRI – Disc extrusion	7.0 (3.3 to 10.7)	98.0 (80.4 to 100.0)	? (?)	? (?)
Ohnmeiss et al 1999	Disc	Pain drawing indicative of disc*	82.3 (74.9 to 88.2)	60.9 (45.4 to 74.9)	2.10 (1.45 to 3.04)	0.29 (0.19 to 0.45)
Ohnmeiss et al 1999	Disc	Aching pain*	78.0 (70.3 to 84.5)	21.7 (10.9 to 36.4)	1.00 (0.84 to 1.19)	1.01 (0.54 to 1.90)
Ohnmeiss et al 1999	Disc	Pins & needles*	32.6 (25.0 to 41.0)	82.6 (68.6 to 92.2)	1.88 (0.96 to 3.68)	0.82 (0.68 to 0.97)
Ohnmeiss et al 1999	Disc	Numbness*	49.6 (41.1 to 58.2)	58.7 (43.2 to 73)	1.20 (0.82 to 1.76)	0.86 (0.64 to 1.15)
Ohnmeiss et al 1999	Disc	Burning*	47.5 (39.1 to 56.1)	78.3 (63.6 to 89.1)	2.19 (1.23 to 3.89)	0.67 (0.54 to 0.84)
Ohnmeiss et al 1999	Disc	Stabbing*	51.1 (42.5 to 59.6)	45.7 (30.9 to 61)	0.94 (0.69 to 1.28)	1.07 (0.75 to 1.53)
Ohnmeiss et al 1999	Disc	Ransford classification normal*	76.6 (68.7 to 83.3)	21.7 (10.9 to 36.4)	0.98 (0.82 to 1.17)	1.08 (0.58 to 2.01)
Osti et al 1992	Disc	MRI- signal intensity decreased & absent	69.2 (52.4 to 83)	64.0 (52.1 to 74.8)	1.92 (1.33 to 2.78)	0.48 (0.29 to 0.79)
Osti et al 1992	Disc	MRI- signal intensity absent	23.1 (11.1 to 39.3)	92.0 (83.4 to 97.0)	2.89 (1.11 to 7.52)	0.84 (0.70 to 1.01)
Parker et al 1996	Disc	MRI-dark disc	90.9 (70.8 to 98.9)	68.8 (41.3 to 89.0)	2.91 (1.39 to 6.09)	0.13 (0.03 to 0.52)
Revel et al 1992	Facet joint	Relief with recumbancy	90.9 (70.8 to 98.9)	44.4 (21.5 to 69.2)	1.64 (1.06 to 2.53)	0.21 (0.05 to 0.85)
Revel et al 1992	Facet joint	Pain not worse with cough	81.8 (59.7 to 94.8)	50.0 (26.0 to 74.0)	1.64 (0.99 to 2.7)	0.36 (0.13 to 0.99)
Revel et al 1992	Facet joint	Pain not worse with forward flexion	63.6 (40.7 to 82.8)	77.8 (52.4 to 93.6)	2.86 (1.14 to 7.19)	0.47 (0.26 to 0.86)
Revel et al 1992	Facet joint	Pain not worse with raising from forward flexion	77.3 (54.6 to 92.2)	55.6 (30.8 to 78.5)	1.74 (0.99 to 3.06)	0.41 (0.17 to 0.98)
Revel et al 1992	Facet joint	Pain not worse with hyperextension	54.5 (32.2 to 75.6)	72.2 (46.5 to 90.3)	1.96 (0.85 to 4.53)	0.63 (0.37 to 1.08)
Revel et al 1992	Facet joint	Pain not worse with extension/rotation	68.2 (45.1 to 86.1)	77.8 (52.4 to 93.6)	3.07 (1.24 to 7.62)	0.41 (0.21 to 0.79)
Revel et al 1992	Facet joint	6 of 7 variables	45.5 (24.4 to 67.8)	97.3 (77.4 to 100)	16.81 (1.05 to 268.9)	0.56 (0.38 to 0.83)
Revel et al 1992	Facet joint	5 of 7 variables	63.6 (40.7 to 82.8)	88.9 (65.3 to 98.6)	5.73 (1.49 to 21.96)	0.41 (0.23 to 0.73)
Revel et al 1992	Facet joint	4 of 7 variables	81.8 (59.7 to 94.8)	77.8 (52.4 to 93.6)	3.68 (1.52 to 8.93)	0.23 (0.09 to 0.59)
Revel et al 1998	Facet joint	5 of 7 characteristics	96.3 (70.3 to 100.0)	65.5 (45.7 to 82.1)	2.79 (1.67 to 4.66)	0.06 (0.00 to 0.87)
Revel et al 1998	Facet joint	5 of 7 characteristics including relieved by recumbancy	92.3 (64.0 to 99.8)	79.3 (60.3 to 92.0)	4.46 (2.15 to 9.26)	0.10 (0.02 to 0.64)
Revel et al 1998	Facet joint	Age > 65	38.5 (13.9 to 68.4)	79.3 (60.3 to 92.0)	1.86 (0.69 to 5.00)	0.78 (0.49 to 1.24)
Revel et al 1998	Facet joint	Pain well relieved by recumbancy	92.3 (64.0 to 99.8)	24.1 (10.3 to 43.5)	1.22 (0.94 to 1.58)	0.32 (0.04 to 2.33)
Revel et al 1998	Facet joint	Pain not exacerbated by coughing	96.3 (70.3 to 100.0)	34.5 (17.9 to 54.3)	1.47 (1.11 to 1.95)	0.11 (0.01 to 1.71)
Revel et al 1998	Facet joint	Pain not exacerbated by forward flexion	96.3 (70.3 to 100.0)	48.3 (29.4 to 67.5)	1.86 (1.29 to 2.69)	0.08 (0.01 to 1.20)
Revel et al 1998	Facet joint	Pain not exacerbated by rising from flexion	96.3 (70.3 to 100.0)	58.6 (38.9 to 76.5)	2.33 (1.49 to 3.63)	0.06 (0.00 to 0.98)
Revel et al 1998	Facet joint	Pain not exacerbated by hyperextension	92.3 (64 to 99.8)	62.1 (42.3 to 79.3)	2.43 (1.49 to 3.98)	0.12 (0.02 to 0.83)
Revel et al 1998	Facet joint	Pain not exacerbated by extension/rotation	76.9 (46.2 to 95)	48.3 (29.4 to 67.5)	1.49 (0.94 to 2.36)	0.48 (0.17 to 1.38)
Ricketson et al 1996	Disc	MRI-HIZ	10.8 (3.0 to 25.4)	93.0 (80.9 to 98.5)	1.55 (0.37 to 6.48)	0.96 (0.83 to 1.10)

Saifuddin et al 1998	Disc	MRI- HIZ	26.7 (17.9 to 37)	95.2 (86.5 to 99)	5.51 (1.74 to 17.5)	0.77 (0.67 to 0.88)
Schellhas et al 1996	Disc	MRI-HIZ	97.8 (92.1 to 99.7)	83.3 (73.2 to 90.8)	5.87 (3.57 to 9.64)	0.03 (0.01 to 0.11)
Schneider et al 2020	SIJ	Thigh thrust	29.4 (10.3 to 56.0)	72.2 (46.5 to 90.3)	1.06 (0.37 to 3.02)	0.98 (0.64 to 1.49)
Schneider et al 2020	SIJ	Gaenslen's test	29.4 (10.3 to 56.0)	44.4 (21.5 to 69.2)	0.53 (0.23 to 1.23)	1.59 (0.87 to 2.90)
Schneider et al 2020	SIJ	FABER test	76.5 (50.1 to 93.2)	27.8 (9.7 to 53.5)	1.06 (0.72 to 1.56)	0.85 (0.27 to 2.64)
Schneider et al 2020	SIJ	Distraction	35.3 (14.2 to 61.7)	83.3 (58.6 to 96.4)	2.12 (0.63 to 7.15)	0.78 (0.52 to 1.17)
Schneider et al 2020	SIJ	Compression	35.3 (14.2 to 61.7)	77.8 (52.4 to 93.6)	1.59 (0.54 to 4.67)	0.83 (0.54 to 1.28)
Schneider et al 2020	SIJ	Sacral thrust	76.5 (50.1 to 93.2)	27.8 (9.7 to 53.5)	1.06 (0.72 to 1.56)	0.85 (0.27 to 2.64)
Schneider et al 2020	SIJ (100% block)	0 test positive	96.0 (68.3 to 100.0)	2.4 (0 to 20.6)	0.98 (0.86 to 1.12)	1.64 (0.04 to 77.55)
Schneider et al 2020	SIJ (100% block)	>1 test positive	96.8 (73.6 to 100.0)	5.0 (0.1 to 24.9)	1.02 (0.89 to 1.17)	0.65 (0.02 to 18.01)
Schneider et al 2020	SIJ (100% block)	>2 test positive	80.0 (51.9 to 95.7)	30.0 (11.9 to 54.3)	1.14 (0.78 to 1.68)	0.67 (0.20 to 2.24)
Schneider et al 2020	SIJ (100% block)	>3 test positive	60.0 (32.3 to 83.7)	45.0 (23.1 to 68.5)	1.09 (0.62 to 1.93)	0.89 (0.41 to 1.95)
Schneider et al 2020	SIJ (100% block)	>4 test positive	33.3 (11.8 to 61.6)	70.0 (45.7 to 88.1)	1.11 (0.42 to 2.96)	0.95 (0.60 to 1.51)
Schneider et al 2020	SIJ (100% block)	>5 test positive	20.0 (4.3 to 48.1)	80.0 (56.3 to 94.3)	1.00 (0.26 to 3.82)	1.00 (0.72 to 1.40)
Schneider et al 2020	SIJ (100% block)	>6 test positive	3.2 (0.0 to 26.4)	90.0 (68.3 to 98.8)	0.32 (0.02 to 6.66)	1.08 (0.91 to 1.28)
Schneider et al 2020	SIJ (80% block)	0 test positive	97.3 (77.4 to 100.0)	2.9 (0.0 to 23.7)	1.00 (0.90 to 1.12)	0.95 (0.02 to 45.12)
Schneider et al 2020	SIJ (80% block)	>1 test positive	97.3 (77.4 to 100.0)	5.9 (0.1 to 28.7)	1.03 (0.90 to 1.19)	0.46 (0.02 to 12.84)
Schneider et al 2020	SIJ (80% block)	>2 test positive	77.8 (52.4 to 93.6)	29.4 (10.3 to 56.0)	1.10 (0.74 to 1.63)	0.76 (0.24 to 2.35)
Schneider et al 2020	SIJ (80% block)	>3 test positive	55.6 (30.8 to 78.5)	41.2 (18.4 to 67.1)	0.94 (0.53 to 1.68)	1.08 (0.50 to 2.33)
Schneider et al 2020	SIJ (80% block)	>4 test positive	33.3 (13.3 to 59)	70.6 (44.0 to 89.7)	1.13 (0.42 to 3.03)	0.94 (0.60 to 1.48)
Schneider et al 2020	SIJ (80% block)	>5 test positive	22.2 (6.4 to 47.6)	82.4 (56.6 to 96.2)	1.26 (0.33 to 4.82)	0.94 (0.68 to 1.32)
Schneider et al 2020	SIJ (80% block)	>6 test positive	5.6 (0.1 to 27.3)	94.1 (71.3 to 99.9)	0.94 (0.06 to 13.93)	1.00 (0.85 to 1.18)
Schneider et al 2020	SIJ (50% block)	0 test positive	98.0 (83.1 to 100.0)	4.8 (0.0 to 36.7)	1.03 (0.89 to 1.19)	0.41 (0.01 to 19.41)
Schneider et al 2020	SIJ (50% block)	>1 test positive	96.0 (79.6 to 99.9)	4.8 (0.0 to 36.7)	1.01 (0.86 to 1.18)	0.84 (0.03 to 23.17)
Schneider et al 2020	SIJ (50% block)	>2 test positive	68.0 (46.5 to 85.1)	10.0 (0.3 to 44.5)	0.76 (0.54 to 1.06)	3.20 (0.46 to 22.38)
Schneider et al 2020	SIJ (50% block)	>3 test positive	48.0 (27.8 to 68.7)	20.0 (2.5 to 55.6)	0.60 (0.36 to 1.00)	2.60 (0.71 to 9.50)
Schneider et al 2020	SIJ (50% block)	>4 test positive	24.0 (9.4 to 45.1)	50.0 (18.7 to 81.3)	0.48 (0.19 to 1.22)	1.52 (0.79 to 2.93)
Schneider et al 2020	SIJ (50% block)	>5 test positive	16.0 (4.5 to 36.1)	70.0 (34.8 to 93.3)	0.53 (0.15 to 1.97)	1.20 (0.77 to 1.86)
Schneider et al 2020	SIJ (50% block)	>6 test positive	4.0 (0.1 to 20.4)	90.0 (55.5 to 99.7)	0.40 (0.03 to 5.79)	1.07 (0.86 to 1.33)
Simmons et al 1991	Disc	MRI abnormal- decreased signal or annular bulge	87.8 (81.9 to 92.3)	63.1 (57.3 to 68.7)	2.38 (2.03 to 2.8)	0.19 (0.13 to 0.29)
Slipman et al 1996	SIJ	Radionuclide imaging (bone scan)	12.9 (3.6 to 29.8)	97.4 (78.4 to 100)	5.03 (0.28 to 90.08)	0.89 (0.77 to 1.04)
Smith et al 1998	Disc	MRI- HIZ vs pain reproduction	22.9 (10.4 to 40.1)	89.7 (82.8 to 94.6)	2.23 (0.99 to 5.02)	0.86 (0.71 to 1.04)
Smith et al 1998	Disc	MRI HIZ vs pain reproduction & morphology	25.8 (11.9 to 44.6)	90.1 (83.3 to 94.8)	2.60 (1.17 to 5.81)	0.82 (0.66 to 1.02)
Stanford et al 2010	SIJ	Patrick's test	? (?)	? (?)	? (?)	? (?)
Stanford et al 2010	SIJ	High thrust test	? (?)	? (?)	? (?)	? (?)
Stanford et al 2010	SIJ	Ipsilateral Gaenslan's test	? (?)	? (?)	? (?)	? (?)
Stanford et al 2010	SIJ	Contralateral Gaenslen's test	? (?)	? (?)	? (?)	? (?)
Stanford et al 2010	SIJ	Lateral compression test	? (?)	? (?)	? (?)	? (?)
Stanford et al 2010	SIJ	Sacral thrust test	? (?)	? (?)	? (?)	? (?)
Stanford et al 2010	SIJ	3 or more positive SIJ provocative tests pre-block	81.8 (48.2 to 97.7)	56.5 (34.5 to 76.8)	1.88 (1.09 to 3.24)	0.57 (0.35 to 0.77)

Stanford et al 2010	SIJ	Normalization of >1/2 of the positive SIJ provocative tests	88.9 (51.8 to 99.7)	30 (6.7 to 65.2)	1.27 (0.8 to 2.03)	0.30 (0.07 to 0.65)
Stojanovic et al 2010	Facet joint	Degneration	77.3 (62.2 to 88.5)	42.3 (31.2 to 54)	1.34 (1.05 to 1.72)	0.54 (0.29 to 0.98)
Stojanovic et al 2010	Facet joint	Hypertrophy	64.1 (47.2 to 78.8)	34.9 (24.8 to 46.2)	0.99 (0.74 to 1.31)	1.03 (0.62 to 1.71)
van der Wurff et al 2006	SIJ	3 or more positive SIJ pain provocation tests	85.2 (66.3 to 95.8)	78.8 (61.1 to 91)	4.02 (2.04 to 7.90)	0.19 (0.08 to 0.47)
Vanharanta et al 1988	Disc	Narrow disc on L4/5 radiograph	26.2 (15.8 to 39.1)	76.1 (61.2 to 87.4)	1.10 (0.56 to 2.13)	0.97 (0.78 to 1.21)
Vanharanta et al 1988	Disc	Narrow disc on L5/S1radiograph	55.2 (42.6 to 67.4)	65.8 (48.6 to 80.4)	1.61 (0.99 to 2.64)	0.68 (0.48 to 0.97)
Vanharanta et al 1988	Disc	Narrow disc height (pooled from 2 levels)	41.4 (32.8 to 50.4)	71.4 (60.5 to 80.8)	1.45 (0.98 to 2.15)	0.82 (0.67 to 1.00)
Vanharanta et al 1988	Disc	Narrowed disc L4/5 radiograph vs pain reproduction & morphology	28.8 (17.1 to 43.1)	78.2 (65 to 88.2)	1.32 (0.69 to 2.55)	0.91 (0.73 to 1.14)
Vanharanta et al 1988	Disc	Narrowed disc L5/S1 radiograph vs pain reproduction & morphology	57.6 (44.1 to 70.4)	65.2 (49.8 to 78.6)	1.66 (1.05 to 2.60)	0.65 (0.45 to 0.94)
Vanharanta et al 1988	Disc	Narrowed disc (pooled L4/5 and L5/S1radiograph) vs pain reproduction & morphology	44.1 (34.7 to 53.9)	72.3 (62.5 to 80.7)	1.59 (1.09 to 2.32)	0.77 (0.63 to 0.95)
Vanharanta et al 1998	Disc	Vibration vs pain reproduction*	70.9 (57.1 to 82.4)	60.9 (38.5 to 80.3)	1.81 (1.06 to 3.10)	0.48 (0.28 to 0.81)
Vanharanta et al 1998	Disc	Vibration vs pain reproduction	60.5 (49.3 to 70.8)	78.3 (69.9 to 85.3)	2.79 (1.91 to 4.08)	0.51 (0.38 to 0.67)
Vanharanta et al 1998	Disc	Vibration Vs pain reproduction and morphology	62.8 (51.1 to 73.5)	77.3 (69.1 to 84.3)	2.77 (1.93 to 3.99)	0.48 (0.36 to 0.65)
Vanharanta et al 1998	Disc	MRI- annular disruption vs pain reproduction & morphology	94.9 (87.4 to 98.6)	54.7 (45.7 to 63.5)	2.09 (1.72 to 2.55)	0.09 (0.04 to 0.25)
Vanharanta et al 1998	Disc	Combined MRI & vibration vs pain reproduction & morphology	85.9 (76.2 to 92.7)	81.3 (73.4 to 87.6)	4.58 (3.16 to 6.64)	0.17 (0.10 to 0.30)
Waldrop et al 2021	Disc	**L3-4**				
Waldrop et al 2021	Disc	Degeneration	91.9 (85.5 to 98.3)	67.2 (63.2 to 71.3)	? (?)	? (?)
Waldrop et al 2021	Disc	Bulge	63.5 (52.3 to 74.7)	86.1 (83.1 to 89.1)	? (?)	? (?)
Waldrop et al 2021	Disc	Tear	40.5 (29.1 to 52.0)	94.3 (92.3 to 96.3)	? (?)	? (?)
Waldrop et al 2021	Disc	Herniation	18.9 (9.8 to 28.1)	98.1 (96.9 to 99.3)	? (?)	? (?)
Waldrop et al 2021	Disc	**L4-5**				
Waldrop et al 2021	Disc	Degeneration	94.3 (91.6 to 97.0)	44.3 (38.6 to 50.0)	? (?)	? (?)
Waldrop et al 2021	Disc	Bulge	73.8 (68.6 to 78.9)	74.5 (69.5 to 79.5)	? (?)	? (?)
Waldrop et al 2021	Disc	Tear	55.0 (49.1 to 60.8)	82.2 (77.9 to 86.6)	? (?)	? (?)
Waldrop et al 2021	Disc	Herniation	27.7 (22.4 to 32.9)	93.3 (90.4 to 96.2)	? (?)	? (?)
Waldrop et al 2021	Disc	**L5-S1**				
Waldrop et al 2021	Disc	Degeneration	96.3 (94.2 to 98.4)	44.1 (38.2 to 50.1)	? (?)	? (?)
Waldrop et al 2021	Disc	Bulge	75.5 (70.7 to 80.2)	75.4 (70.2 to 80.5)	? (?)	? (?)
Waldrop et al 2021	Disc	Tear	68.6 (63.5 to 73.7)	86.8 (82.7 to 90.8)	? (?)	? (?)
Waldrop et al 2021	Disc	Herniation	42.9 (37.4 to 48.3)	93.8 (90.9 to 96.6)	? (?)	? (?)
Weishaupt et al 2001	Disc	MRI -disc degeneration	97.9 (88.9 to 99.9)	58.8 (46.2 to 70.6)	2.38 (1.79 to 3.17)	0.04 (0.01 to 0.25)
Weishaupt et al 2001	Disc	MRI- HIZ	27.1 (15.3 to 41.8)	85.3 (74.6 to 92.7)	1.84 (0.88 to 3.85)	0.86 (0.70 to 1.04)
Weishaupt et al 2001	Disc	MRI- Modic changes grade 1 (all)	29.2 (17 to 44.1)	97.1 (89.8 to 99.6)	9.92 (2.36 to 41.63)	0.73 (0.61 to 0.88)
Weishaupt et al 2001	Disc	MRI- Modic changes grade 1(mod and severe)	22.9 (12 to 37.3)	99.3 (93.4 to 100)	31.39 (1.89 to 521.4)	0.78 (0.67 to 0.91)
Weishaupt et al 2001	Disc	MRI- Modic changes grade 2 (all)	18.8 (8.9 to 32.6)	98.5 (92.1 to 100)	12.75 (1.67 to 97.34)	0.83 (0.72 to 0.95)
Weishaupt et al 2001	Disc	MRI- Modic changes grade 2 (mod and severe)	14.6 (6.1 to 27.8)	99.3 (93.4 to 100)	19.97 (1.16 to 343.7)	0.86 (0.76 to 0.97)

Weishaupt et al 2001	Disc	MRI- Modic changes grades 1&2 (all)	47.9 (33.3 to 62.8)	95.6 (87.6 to 99.1)	10.86 (3.46 to 34.13)	0.55 (0.41 to 0.72)
Weishaupt et al 2001	Disc	MRI- Modic changes grades 1 &2 (mod and severe)	37.5 (24 to 52.6)	99.3 (93.4 to 100)	51.37 (3.17 to 832.8)	0.63 (0.51 to 0.79)
Yoshida et al 2002	Disc	MRI - post annulus tear on T2 scan	94.1 (71.3 to 99.9)	71.8 (55.1 to 85)	3.34 (2.00 to 5.58)	0.08 (0.01 to 0.55)
Yoshida et al 2002	Disc	MRI - post annulus tear on Gd-DTPA-enhanced T1 scan	71.4 (41.9 to 91.6)	75 (56.6 to 88.5)	2.86 (1.44 to 5.67)	0.38 (0.16 to 0.89)
Young et al 2003	SIJ	Absense of midline lumbar pain	38.5 (20.2 to 59.4)	26.7 (12.3 to 45.9)	2.31 (1.19 to 4.50)	0.52 (0.31 to 0.89)
Young et al 2003	SIJ	Asymmetrical pain	26.9 (11.6 to 47.8)	53.3 (34.3 to 71.7)	0.58 (0.28 to 1.21)	1.37 (0.91 to 2.06)
Young et al 2003	SIJ	Unilateral pain	73.1 (52.2 to 88.4)	38.7 (21.8 to 57.8)	1.19 (0.83 to 1.72)	0.70 (0.32 to 1.51)
Young et al 2003	SIJ	Bilateral pain	26.9 (11.6 to 47.8)	63.3 (43.9 to 80.1)	0.73 (0.33 to 1.62)	1.15 (0.81 to 1.65)
Young et al 2003	SIJ	No change in lumbar range of motion	98.1 (83.7 to 100)	13.3 (3.8 to 30.7)	1.13 (0.97 to 1.32)	0.14 (0.01 to 2.55)
Young et al 2003	SIJ	No Centralisation	92.3 (74.9 to 99.1)	23.3 (9.9 to 42.3)	1.20 (0.96 to 1.51)	0.33 (0.08 to 1.45)
Young et al 2003	SIJ	No Peripheralisation	96.2 (80.4 to 99.9)	23.3 (9.9 to 42.3)	1.25 (1.02 to 1.55)	0.17 (0.02 to 1.25)
Young et al 2003	SIJ	Pain with rising	97.9 (81.8 to 100)	26.7 (12.3 to 45.9)	1.34 (1.07 to 1.67)	0.08 (0.01 to 1.32)
Young et al 2003	SIJ	3 or more positive SIJ pain provocation test	76.9 (56.4 to 91)	70 (50.6 to 85.3)	2.56 (1.43 to 4.61)	0.33 (0.16 to 0.69)
Young et al 2003	Disc	Midline lumbar pain*	26.7 (7.8 to 55.1)	37.5 (8.5 to 75.5)	0.43 (0.16 to 1.16)	1.96 (0.76 to 5.03)
Young et al 2003	Disc	Asymmetrical pain*	60 (32.3 to 83.7)	66.7 (29.9 to 92.5)	1.80 (0.65 to 4.95)	0.60 (0.28 to 1.30)
Young et al 2003	Disc	Unilateral pain*	33.3 (11.8 to 61.6)	55.6 (21.2 to 86.3)	0.75 (0.27 to 2.09)	1.20 (0.61 to 2.38)
Young et al 2003	Disc	Bilateral pain*	40 (16.3 to 67.7)	55.6 (21.2 to 86.3)	0.90 (0.35 to 2.35)	1.08 (0.53 to 2.21)
Young et al 2003	Disc	Change in lumbar range of motion*	13.3 (1.7 to 40.5)	94.7 (60.3 to 100)	2.53 (0.13 to 50.4)	0.92 (0.71 to 1.17)
Young et al 2003	Disc	Centralisation*	46.7 (21.3 to 73.4)	94.7 (60.3 to 100)	8.87 (0.57 to 138.90)	0.56 (0.34 to 0.93)
Young et al 2003	Disc	Peripheralisation*	26.7 (7.8 to 55.1)	94.7 (60.3 to 100)	5.07 (0.30 to 85.46)	0.77 (0.55 to 1.09)
Young et al 2003	Disc	Pain with rising*	96.8 (73.6 to 100)	33.3 (7.5 to 70.1)	1.45 (0.91 to 2.32)	0.10 (0.01 to 1.72)
Young et al 2003	Disc	Less than 3 positive SIJ pain provocation tests*	72.7 (39 to 94)	50 (15.7 to 84.3)	1.46 (0.67 to 3.18)	0.55 (0.17 to 1.79)
Young et al 2003	Facet joint	Non midline lumbar pain	76.9 (46.2 to 95)	22.2 (2.8 to 60)	0.99 (0.63 to 1.57)	1.04 (0.22 to 5.01)
Young et al 2003	Facet joint	Asymmetrical pain	35.7 (12.8 to 64.9)	66.7 (29.9 to 92.5)	1.07 (0.34 to 3.42)	0.96 (0.53 to 1.77)
Young et al 2003	Facet joint	Unilateral pain	71.4 (41.9 to 91.6)	44.4 (13.7 to 78.8)	1.29 (0.66 to 2.52)	0.64 (0.21 to 1.94)
Young et al 2003	Facet joint	Non-Bilateral pain	92.9 (66.1 to 99.8)	22.2 (2.8 to 60)	1.19 (0.82 to 1.74)	0.32 (0.03 to 3.05)
Young et al 2003	Facet joint	No change in lumbar range of motion	96.6 (72 to 100)	5.3 (0 to 39.7)	1.02 (0.85 to 1.22)	0.66 (0.01 to 30.28)
Young et al 2003	Facet joint	No centralisation	96.6 (72 to 100)	11.1 (0.3 to 48.2)	1.09 (0.85 to 1.40)	0.31 (0.01 to 8.34)
Young et al 2003	Facet joint	No peripheralisation	96.6 (72 to 100)	11.1 (0.3 to 48.2)	1.09 (0.85 to 1.40)	0.31 (0.01 to 8.34)
Young et al 2003	Facet joint	No pain with rising	78.6 (49.2 to 95.3)	77.8 (40.0 to 97.2)	3.54 (1.01 to 12.37)	0.28 (0.10 to 0.80)
Young et al 2003	Facet joint	Less than 3 positive SIJ tests	66.7 (34.9 to 90.1)	22.2 (2.8 to 60.0)	0.86 (0.50 to 1.46)	1.50 (0.35 to 6.47)
Yrjama et al 1994	Disc	Vibration *	71.1 (54.1 to 84.6)	63.2 (38.4 to 83.7)	1.93 (1.04 to 3.60)	0.46 (0.25 to 0.84)
Yrjama et al 1994	Disc	vibration	81 (58.1 to 94.6)	84.6 (54.6 to 98.1)	5.26 (1.45 to 19.14)	0.23 (0.09 to 0.56)
Yrjama et al 1996	Disc	Ultrasound grade 3*	61.5 (40.6 to 79.8)	33.3 (9.9 to 65.1)	0.92 (0.56 to 1.53)	1.15 (0.45 to 2.94)
Yrjama et al 1996	Disc	Vibration *	65.4 (44.3 to 82.8)	58.3 (27.7 to 84.8)	1.57 (0.76 to 3.24)	0.59 (0.29 to 1.21)
Yrjama et al 1997	Disc	MRI- annular disruption ≥ grade 3	65.2 (42.7 to 83.6)	66.7 (22.3 to 95.7)	1.96 (0.61 to 6.31)	0.52 (0.24 to 1.16)
Yrjama et al 1997	Disc	Vibration	60.9 (38.5 to 80.3)	66.7 (22.3 to 95.7)	1.83 (0.56 to 5.93)	0.59 (0.27 to 1.26)

SIJ=sacroiliac joint; MRI=magnetic resonance imaging; CT=computed tomography; SPECT=single-proton emission computed tomography; FABER=flexion abduction external rotation test; HIZ=high intensity zone; SLR=straight leg raise
*Analysis at patient level, all other studies on the disc are based on the disc level

Appendix S7. Pooled diagnostic values of index tests

Index test	Studies/Total sample	Sensitivity (95%CI)	Specificity (95%CI)	LR+ (95%CI)	LR- (95%CI)
Disc					
Disc degeneration (Grade≥3)[34]	116	97.9 (88.9 to 99.9)	58.8 (46.2 to 70.6)	2.38 (1.79 to 3.17)	0.04 (0.01 to 0.25)
Disc degeneration (Grade≥3)[4]	152	86.7 (77.9 to 92.9)	79 (66.8 to 88.3)	4.13 (2.53 to 6.75)	0.17 (0.10 to 0.29)
Disc degeneration (Grade≥3)[24]	38	90.9 (70.8 to 98.9)	68.8 (41.3 to 89)	2.91 (1.39 to 6.09)	0.13 (0.03 to 0.52)
Disc degeneration (Grade≥3)[5]	75	94.1 (71.3 to 99.9)	43.1 (30.2 to 56.8)	1.65 (1.28 to 2.13)	0.14 (0.02 to 0.94)
Pooled result					
Disc degeneration (Grade≥3)[4,5,24,34]	**4/381**	**91.0 (85.7 to 94.7)**	**61.3 (54.2 to 68.0)**	**2.53 (1.57 to 4.07)**	**0.15 (0.09 to 0.24)**
Heterogeneity (I^2)		***50.1%***	***82.7%***	***83.4%***	***0.0%***
Disc degeneration (Grade≥4)*[34]	116	54.2 (39.2 to 68.6)	83.8 (72.9 to 91.6)	3.35 (1.84 to 6.10)	0.55 (0.40 to 0.76)
Disc degeneration (Grade≥4)*[19]	97	88.2 (72.5 to 96.7)	52.4 (39.4 to 65.1)	1.85 (1.39 to 2.47)	0.23 (0.09 to 0.58)
Disc degeneration (Grade≥4)*[5]	75	82.4 (56.6 to 96.2)	62.1 (48.4 to 74.5)	2.17 (1.46 to 3.23)	0.28 (0.10 to 0.81)
Pooled result					
Disc degeneration (Grade≥4)*[5,19,34]	**3/288**	**70.7 (60.7 to 79.4)**	**66.7 (59.5 to 73.3)**	**2.20 (1.61 to 3.01)**	**0.37 (0.19 to 0.73)**
Heterogeneity (I^2)		***84.7%***	***87.7%***	***44.4%***	***59.9%***
HIZ[18]	113	24.5 (13.8 to 38.3)	86.7 (75.4 to 94.1)	1.84 (0.83 to 4.09)	0.87 (0.75 to 0.94)
HIZ[10]	119	43.7 (31.9 to 56)	72.9 (58.2 to 84.7)	1.61 (0.95 to 2.75)	0.73 (0.58 to 0.85)
HIZ[25]	80	10.8 (3.0 to 25.4)	93.0 (80.9 to 98.5)	1.55 (0.37 to 6.48)	0.96 (0.83 to 1.10)
HIZ[27]	167	97.8 (92.1 to 99.7)	83.3 (73.2 to 90.8)	5.87 (3.57 to 9.64)	0.03 (0.01 to 0.11)
HIZ[19]	97	55.9 (37.9 to 72.8)	69.8 (57.0 to 80.8)	1.85 (1.15 to 3.00)	0.63 (0.42 to 0.95)
HIZ[34]	116	27.1 (15.3 to 41.8)	85.3 (74.6 to 92.7)	1.84 (0.88 to 3.85)	0.86 (0.70 to 1.04)
HIZ[12]	101	52.2 (30.6 to 73.2)	89.7 (80.8 to 95.5)	5.09 (2.37 to 10.92)	0.53 (0.35 to 0.82)
HIZ[26]	152	26.7 (17.9 to 37)	95.2 (86.5 to 99)	5.51 (1.74 to 17.5)	0.77 (0.67 to 0.88)
HIZ[23]	460	44.4 (38.0 to 50.9)	89.1 (84.3 to 92.9)	4.08 (2.73 to 6.11)	0.62 (0.55 to 0.71)
HIZ[2]	105	70.6 (56.2 to 82.5)	88.9 (77.4 to 95.8)	6.35 (2.93 to 13.78)	0.33 (0.21 to 0.52)
HIZ[13]	155	80.8 (71.7 to 88.0)	78.6 (65.6 to 88.4)	3.77 (2.26 to 6.28)	0.24 (0.16 to 0.37)
HIZ[30]	152	25.8 (11.9 to 44.6)	90.1 (83.3 to 94.8)	2.60 (1.17 to 5.81)	0.82 (0.66 to 1.02)
Pooled result					
HIZ[2,10,12,13,18,19,23,25-27,30,34]	**12/1817**	**50.1 (46.7 to 53.4)**	**86.2 (83.9 to 88.4)**	**3.10 (2.27 to 4.25)**	**0.61 (0.48 to 0.77)**
Heterogeneity (I^2)		***95.4%***	***65.6%***	***63.6%***	***92.2%***
Annular fissure[12]	99	95.2 (76.2 to 99.9)	55.1 (43.4 to 66.4)	2.12 (1.63 to 2.76)	0.09 (0.01 to 0.59)
Annular fissure[35]	56	94.1 (71.3 to 99.9)	71.8 (55.1 to 85)	3.34 (1.99 to 5.58)	0.08 (0.01 to 0.55)
Annular fissure[33]	206	94.9 (87.4 to 98.6)	54.7 (45.7 to 63.5)	2.09 (1.72 to 2.55)	0.09 (0.04 to 0.25)
Annular fissure[23]	460	44.4 (38.0 to 50.9)	89.1 (84.3 to 92.9)	4.08 (2.73 to 6.11)	0.62 (0.55 to 0.71)
Annular fissure[3]	99	64.0 (49.2 to 77.1)	85.7 (72.8 to 94.1)	4.48 (2.19 to 9.17)	0.42 (0.29 to 0.62)
Pooled result					
Annular fissure[3,12,23,33,35]	**5/920**	**61.2 (56.3 to 66.0)**	**73.8 (69.8 to 77.5)**	**2.88 (2.02 to 4.10)**	**0.24 (0.10 to 0.55)**
Heterogeneity (I^2)		***96.0%***	***94.3%***	***79.5%***	***90.3%***
Modic type 1[23]	440	6.8 (3.9 to 11.0)	99.1 (96.7 to 99.9)	7.43 (1.72 to 32.10)	0.94 (0.91 to 0.98)
Modic type 1[4]	134	6.8 (2.2 to 15.1)	99.2 (92.5 to 100.0)	8.18 (0.46 to 146.70)	0.94 (0.88 to 1.00)
Modic type 1[34]	116	29.2 (17 to 44.1)	97.1 (89.8 to 99.6)	9.92 (2.36 to 41.63)	0.73 (0.61 to 0.88)
Modic type 1[18]	113	32.1 (19.9 to 46.3)	98.3 (91.1 to 100)	19.24 (2.65 to 139.76)	0.98 (0.91 to 1.00)
Pooled result					
Modic type 1[4,18,23,34]	**4/803**	**12.9 (9.7 to 16.6)**	**98.7 (97.0 to 99.5)**	**10.0 (4.20 to 23.82)**	**0.84 (0.74 to 0.96)**
Heterogeneity (I^2)		***91.0%***	***0.0%***	***0.0%***	***89.2%***
Modic type 2[23]	443	6.8 (3.9 to 11.0)	99.1 (96.7 to 99.9)	7.43 (1.72 to 32.10)	0.94 (0.91 to 0.98)
Modic type 2[4]	147	6.8 (2.2 to 15.1)	99.2 (92.5 to 100.0)	8.18 (0.46 to 146.70)	0.94 (0.88 to 1.00)
Modic type 2[34]	116	29.2 (17 to 44.1)	97.1 (89.8 to 99.6)	9.92 (2.36 to 41.63)	0.73 (0.61 to 0.88)
Pooled result					
Modic type 2[4,23,34]	**3/706**	**12.0 (8.9 to 15.9)**	**98.6 (96.7 to 99.5)**	**8.03 (3.23 to 19.97)**	**0.88 (0.80 to 0.96)**
Heterogeneity (I^2)		***77.3%***	***0.0%***	***0.0%***	***65.8%***
†Centralisation phenomenon[7]	63	63.9 (46.2 to 79.2)	70.4 (49.8 to 86.2)	2.16 (1.15 to 4.05)	0.51 (0.31 to 0.85)
†Centralisation phenomenon[36]	24	46.7 (21.3 to 73.4)	94.7 (60.3 to 100.0)	8.87 (0.57 to 138.9)	0.56 (0.34 to 0.93)
†Centralisation phenomenon[16]	69	40.4 (27.0 to 54.9)	94.1 (71.3 to 99.9)	6.87 (1.00 to 47.29)	0.63 (0.49 to 0.82)
†Centralisation phenomenon[14]	62	22.2 (11.2 to 37.1)	97.1 (76.3 to 100)	7.78 (0.48 to 126.1)	0.80 (0.67 to 0.95)
Pooled result					
†Centralisation phenomenon[7,14,16,36]	**4/218**	**41.2 (33.2 to 49.6)**	**85.9 (75.6 to 93.0)**	**3.06 (1.44 to 6.50)**	**0.66 (0.52 to 0.84)**
Heterogeneity (I^2)		***79.9%***	***66.1%***	***10.8%***	***57.9%***
SIJ					
Radionuclide imaging (bone scan)[20]	32	46.2 (19.2 to 74.9)	94.7 (74 to 99.9)	8.77 (1.19 to 64.53)	0.57 (0.34 to 0.95)
Radionuclide imaging (bone scan)[29]	50	12.9 (3.6 to 29.8)	97.4 (78.4 to 100)	5.03 (0.28 to 90.08)	0.89 (0.77 to 1.04)
Pooled result					
Radionuclide imaging (bone scan)[20,29]	**2/82**	**22.7 (11.5 to 37.8)**	**96.1 (84.4 to 99.7)**	**7.33 (1.42 to 37.8)**	**0.74 (0.41 to 1.34)**
Heterogeneity (I^2)		***81.4%***	***0.0%***	***0.0%***	***80.2%***
†Gaenslen's test[15]	45	38.9 (17.3 to 64.3)	74.1 (53.7 to 88.9)	1.5 (0.63 to 3.55)	0.83 (0.54 to 1.27)
†Gaenslen's test[8]	85	66.7 (51.0 to 80.0)	35 (20.6 to 51.7)	1.03 (0.75 to 1.40)	0.95 (0.53 to 1.72)
†Gaenslen's test[22]	48	38.5 (23.4 to 55.4)	33.3 (7.5 to 70.1)	0.58 (0.31 to 1.06)	1.85 (0.71 to 4.81)
†Gaenslen's test[28]	35	29.4 (10.3 to 56.0)	44.4 (21.5 to 69.2)	0.53 (0.23 to 1.23)	1.59 (0.87 to 2.90)
Pooled result					
†Gaenslen's test[8,15,22,28]	**4/213**	**47.9 (38.7 to 57.2)**	**47.9 (37.5 to 58.4)**	**0.85 (0.56 to 1.28)**	**1.12 (0.77 to 1.62)**

Heterogeneity (I^2)		***72.3%***	***73.3%***	***46.5%***	***34.3%***
†Sacral thrust test[8]	85	51.1 (35.8 to 66.3)	40.0 (24.9 to 56.7)	0.85 (0.58 to 1.25)	1.22 (0.75 to 1.98)
†Sacral thrust test[15]	48	55.0 (31.5 to 76.9)	75.0 (55.1 to 89.3)	2.20 (1.04 to 4.68)	0.60 (0.35 to 1.02)
†Sacral thrust test[28]	35	76.5 (50.1 to 93.2)	27.8 (9.7 to 53.5)	1.06 (0.72 to 1.56)	0.85 (0.27 to 2.64)
Pooled result					
†Sacral thrust test[8,15,28]	**3/168**	**57.3 (45.9 to 68.2)**	**48.8 (37.9 to 59.9)**	**1.13 (0.73 to 1.75)**	**0.87 (0.52 to 1.44)**
Heterogeneity (I^2)		***42.5%***	***84.1%***	***59.7%***	***47.7%***
†Thigh thrust test[8]	85	42.2 (27.7 to 57.8)	45 (29.3 to 61.5)	0.77 (0.49 to 1.19)	1.28 (0.84 to 1.96)
†Thigh thrust test[15]	48	70 (45.7 to 88.1)	64.3 (44.1 to 81.4)	1.96 (1.1 to 3.48)	0.47 (0.23 to 0.96)
†Thigh thrust test[22]	48	74.4 (57.9 to 87.0)	44.4 (13.7 to 78.8)	1.34 (0.73 to 2.47)	0.58 (0.23 to 1.43)
†Thigh thrust test[21]	199	53.2 (45.1 to 61.1)	48.8 (32.9 to 64.9)	1.04 (0.74 to 1.45)	0.96 (0.67 to 1.37)
†Thigh thrust test[28]	35	29.4 (10.3 to 56.0)	72.2 (46.5 to 90.3)	1.06 (0.37 to 3.02)	0.98 (0.64 to 1.49)
Pooled result					
†Thigh thrust test[8,15,21,22,28]	**5/415**	**54.1 (48.1 to 60.1)**	**53.7 (44.9 to 62.3)**	**1.13 (0.83 to 1.55)**	**0.91 (0.67 to 1.22)**
Heterogeneity (I^2)		***74.6%***	***31.0%***	***42.5%***	***42.1%***
†Compression test[15]	48	60 (36.1 to 80.9)	67.9 (47.6 to 84.1)	1.87 (0.98 to 3.56)	0.59 (0.33 to 1.07)
†Compression test[28]	35	35.3 (14.2 to 61.7)	77.8 (52.4 to 93.6)	1.59 (0.54 to 4.67)	0.83 (0.54 to 1.28)
Pooled result					
†Compression test[15,28]	**2/83**	**48.6 (31.9 to 65.6)**	**71.7 (56.5 to 84.0)**	**1.79 (1.03 to 3.11)**	**0.74 (0.52 to 1.05)**
Heterogeneity (I^2)		***56.0%***	***0.0%***	***0.0%***	***0.0%***
†Patrick's test (FABER test)[8]	85	57.8 (42.2 to 72.3)	22.5 (10.8 to 38.5)	0.75 (0.55 to 1.01)	1.88 (0.96 to 3.66)
†Patrick's test (FABER test)[21]	199	81.6 (74.7 to 87.3)	43.9 (28.5 to 60.3)	1.46 (1.10 to 1.93)	0.42 (0.26 to 0.67)
†Patrick's test (FABER test)[28]	35	76.5 (50.1 to 93.2)	27.8 (9.7 to 53.5)	1.06 (0.72 to 1.56)	0.85 (0.27 to 2.64)
Pooled result					
†Patrick's test (FABER test)[8,21,28]	**2/319**	**76.4 (70.2 to 81.8)**	**32.3 (23.3 to 42.5)**	**1.05 (0.69 to 1.60)**	**0.86 (0.30 to 2.48)**
Heterogeneity (I^2)		***80.3%***	***55.0%***	***80.7%***	***84.9%***
†Distraction test[15]	47	47.4 (24.4 to 71.1)	78.6 (59 to 91.7)	2.21 (0.94 to 5.19)	0.67 (0.42 to 1.07)
†Distraction test[28]	35	35.3 (14.2 to 61.7)	83.3 (58.6 to 96.4)	2.12 (0.63 to 7.15)	0.78 (0.52 to 1.17)
Pooled result					
†Distraction test[15,28]	**2/82**	**41.7 (25.5 to 59.2)**	**80.4 (66.1 to 90.6)**	**2.18 (1.08 to 4.38)**	**0.73 (0.54 to 0.99)**
Heterogeneity (I^2)		***0.0%***	***0.0%***	***0.0%***	***0.0%***
†Gillet's test[8]	85	40.0 (25.7 to 55.7)	55.0 (38.5 to 70.7)	0.89 (0.54 to 1.46)	1.09 (0.76 to 1.58)
†Gillet's test[22]	49	98.7 (88.8 to 100.0)	5.3 (0.00 to 39.7)	1.04 (0.89 to 1.22)	0.24 (0.50 to 11.36)
Pooled result					
†Gillet's test[8,22]	**2/134**	**67.5 (56.4 to 77.3)**	**45.5 (31.2 to 60.2)**	**1.01 (0.80 to 1.28)**	**1.08 (0.75 to 1.55)**
Heterogeneity (I^2)		***97.5%***	***89.2%***	***23.8%***	***0.0%***
†Absence of midline LBP[6]	170	12.9 (3.6 to 29.8)	36.0 (28.0 to 44.5)	2.42 (1.87 to 3.14)	0.20 (0.08 to 0.51)
†Absence of midline LBP[36]	56	38.5 (20.2 to 59.4)	26.7 (12.3 to 45.9)	2.31 (1.19 to 4.50)	0.52 (0.31 to 0.89)
Pooled result					
†Absence of midline LBP[6,36]	**2/226**	**24.6 (14.1 to 37.8)**	**34.3 (27.2 to 42.0)**	**2.41 (1.89 to 3.07)**	**0.35 (0.12 to 1.01)**
Heterogeneity (I^2)		***80.2%***	***0.0%***	***0.0%***	***75.6%***
†3 or more positive SIJ pain provocation tests[36]	56	76.9 (56.4 to 91.0)	70.0 (50.6 to 85.3)	2.56 (1.43 to 4.61)	0.33 (0.16 to 0.69)
†3 or more positive SIJ pain provocation tests[15]	48	75.0 (50.9 to 91.3)	70.0 (55.1 to 89.3)	3.00 (1.51 to 5.98)	0.33 (0.15 to 0.73)
†3 or more positive SIJ pain provocation tests[17]	43	90.9 (58.7 to 99.8)	78.1 (60.0 to 90.7)	4.16 (2.10 to 8.21)	0.12 (0.02 to 0.76)
†3 or more positive SIJ pain provocation tests[32]	60	85.2 (66.3 to 95.8)	78.8 (61.1 to 91)	4.02 (2.04 to 7.90)	0.19 (0.08 to 0.47)
†3 or more positive SIJ pain provocation tests[31]	34	81.8 (48.2 to 97.7)	56.5 (34.5 to 76.8)	1.88 (1.09 to 3.24)	0.57 (0.35 to 0.77)
†3 or more positive SIJ pain provocation tests[28]	35	77.8 (52.4 to 93.6)	29.4 (10.3 to 56.0)	1.10 (0.74 to 1.63)	0.76 (0.24 to 2.35)
Pooled result					
†3 or more positive SIJ pain provocation tests[15,17,28,31,32,36]	**6/276**	**80.5 (72.0 to 87.4)**	**68.1 (60.4 to 75.2)**	**2.44 (1.50 to 3.98)**	**0.31 (0.21 to 0.47)**
Heterogeneity (I^2)		***0.0%***	***69.1%***	***76.6%***	***0.0%***
Facet joint					
Uptake on SPECT[1]	51	71.4 (55.4 to 84.3)	66.7 (29.9 to 92.5)	2.14 (0.83 to 5.51)	0.43 (0.22 to 0.83)
Uptake on SPECT[9]	29	57.1 (18.4 to 90.1)	77.3 (54.6 to 92.2)	2.51 (0.92 to 6.85)	0.56 (0.23 to 1.34)
Uptake on SPECT[11]	41	93.3 (52.4 to 100.0)	70.6 (52.5 to 84.9)	3.17 (1.82 to 5.53)	0.94 (0.01 to 1.39)
Pooled result					
Uptake on SPECT	**3/121**	**72.6 (59.1 to 83.6)**	**72.3 (59.8 to 82.7)**	**2.80 (1.82 to 4.31)**	**0.44 (0.25 to 0.77)**
Heterogeneity (I^2)		***31.1%***	***0.0%***	***0.0%***	***7.0%***

†All index tests are based on MRI unless indicated

*MRI finding index compared to disc degeneration (Grade ≤ 3), SIJ=Sacroiliac joint, LR=likelihood ratio, SPECT= SPECT=single photon emission computed tomography

All MRI findings (index tests) are compared to the absence of the respected MRI finding unless indicated

No index tests based on imaging were able to be pooled for Facet joint

Appendix S8. Diagnostic values of sensitivity analysis

Index tests	Studies/Total sample	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95%CI)	LR- (95%CI)
Pooling all studies regardless of reference standard quality (i.e., stricter reference standard)					
HIZ[2,10,12,13,18,19,23,25-27,30,34]					
Pooled result	12/1817	50.1 (46.7 to 53.4)	86.2 (83.9 to 88.4)	3.10 (2.27 to 4.25)	0.61 (0.48 to 0.77)
3 or more positive SIJ pain provocation tests[15,17,31,32,36]					
Pooled result	5/241	81.1 (71.7 to 88.4)	72.6 (64.6 to 79.7)	2.87 (2.11 to 3.88)	0.28 (0.18 to 0.43)
Pooling based on stricter reference standard quality					
HIZ[2,10,23,25,27] (Controlled disc)					
Pooled result	5/886	54.2 (49.7 to 58.7)	86.7 (83.2 to 89.7)	3.54 (2.03 to 6.20)	0.48 (0.28 to 0.81)
3 or more positive SIJ pain provocation tests[17,32] (Double injection)					
Pooled result	2/103	86.8 (71.9 to 95.6)	78.5 (66.5 to 87.7)	4.09 (2.53 to 6.60)	0.17 (0.08 to 0.39)
Pooling all studies regardless of risk of bias score					
HIZ[2,10,12,13,18,19,23,25-27,30,34]					
Pooled result	12/1817	50.1 (46.7 to 53.4)	86.2 (83.9 to 88.4)	3.10 (2.27 to 4.25)	0.61 (0.48 to 0.77)
Heterogeneity (I^2)		*95.4%*	*65.6%*	*63.6%*	*92.2%*
3 or more positive SIJ pain provocation tests[15,17,31,32,36]					
Pooled result	5/241	81.1 (71.7 to 88.4)	72.6 (64.6 to 79.7)	2.87 (2.11 to 3.88)	0.28 (0.18 to 0.43)
Heterogeneity (I^2)		*0.0%*	*69.1%*	*76.6%*	*0.0%*
Pooling based on high risk of bias removed for domain 'patient selection'					
HIZ[2,10,12,18,23,25,27]					
Pooled result	7/1145	51.3 (47.2 to 55.5)	87.1 (84.1 to 89.7)	3.43 (2.19 to 5.37)	0.56 (0.38 to 0.81)
Heterogeneity (I^2)		*96.2%*	*44.8%*	*69.8%*	*94.5%*
Pooling based on high risk of bias removed for domain 'index test'					
HIZ[2,10,12,13,18,19,23,25,26,30,34]					
Pooled result	11/1650	44.6 (41.1 to 48.2)	86.5 (84.1 to 88.7)	2.88 (2.11 to 3.93)	0.66 (0.55 to 0.80)
Heterogeneity (I^2)		*91.9%*	*68.2%*	*56.9%*	*88.4%*
Pooling based on high risk of bias removed for domain 'reference standard'					
HIZ[19,25-27,30,34]					
Pooled result	6/764	47.1 (41.6 to 52.7)	86.2 (82.6 to 89.3)	2.83 (1.65 to 4.84)	0.68 (0.47 to 0.97)
Heterogeneity (I^2)		*97.0%*	*76.0%*	*67.9%*	*93.7%*
Pooling based on high risk of bias removed for domain 'flow and timing'					
HIZ[2,12,18,19,23,26,30,34]					
Pooled result	8/1296	40.6 (36.5 to 44.8)	87.6 (85.0 to 89.9)	3.12 (2.15 to 4.52)	0.69 (0.59 to 0.82)
Heterogeneity (I^2)		*84.5%*	*65.0%*	*57.5%*	*78.9%*
Pooling based on high risk of bias removed for domain 'patient selection'					
3 or more positive SIJ pain provocation tests[28,31,36]					
Pooled result	3/125	78.2 (65.0 to 88.2)	55.7 (43.3 to 67.6)	1.69 (0.98 to 2.90)	0.40 (0.23 to 0.70)
Heterogeneity (I^2)		0%	72.90%	71.10%	0.00%
Pooling based on high risk of bias removed for domain 'index test'					
3 or more positive SIJ pain provocation tests[15,17,31,32,36]					
Pooled result	5/241	81.1 (71.7 to 88.4)	72.6 (64.6 to 79.7)	2.87 (2.11 to 3.88)	0.28 (0.18 to 0.43)
Heterogeneity (I^2)		0.00%	2.30%	14.60%	0.00%
Pooling based on high risk of bias removed for domain 'reference standard'					
3 or more positive SIJ pain provocation tests[17,31,32]					
Pooled result	3/137	85.7 (72.8 to 94.1)	72.7 (62.2 to 81.7)	3.06 (1.75 to 5.33)	0.21 (0.10 to 0.41)
Heterogeneity (I^2)		*0.0%*	*48.8%*	*57.2%*	*0.0%*
Pooling based on high risk of bias removed for domain 'flow and timing'					
3 or more positive SIJ pain provocation tests[15,28,31,32]					
Pooled result	4/177	80.3 (69.5 to 88.5)	64.4 (54.2 to 73.6)	2.15 (1.12 to 4.12)	0.33 (0.19 to 0.57)
Heterogeneity (I^2)		0.0%	78.4%	81.4%	15.2%

HIZ=high intensity zone; SIJ=sacroiliac joint; LR=likelihood ratio

MRI findings (index tests) are compared to the absence of the respected MRI finding unless indicated

Reference standard for disc = controlled pain free disc; reference standard for SIJ = double block; N/A = not applicable as all studies were of low risk of bias in that domain

References

1 Akgun E, Akgun MY. The effectiveness of bone scintigraphy in the management of low back pain. Clin Neurol Neurosurg 2022;222:107440.

2 Aprill C, Bogduk N. High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging. Br J Radiol 1992;65(773):361-369.

3 Bartynski WS, Agarwal V, Trang H, Bandos AI, Rothfus WE, Tsay J, Delfyett WT, Nastasi B. Enhancing Annular Fissures and High-Intensity Zones: Pain, Internal Derangement, and Anesthetic Response at Provocation Lumbar Discography. AJNR Am J Neuroradiol 2023;44(1):95-104.

4 Braithwaite I, White J, Saifuddin A, Renton P, Taylor BA. Vertebral end-plate (Modic) changes on lumbar spine MRI: correlation with pain reproduction at lumbar discography. Eur Spine J 1998;7(5):363-368.

5 Carragee EJ, Alamin TF, Miller J, Grafe M. Provocative discography in volunteer subjects with mild persistent low back pain. Spine J 2002;2(1):25-34.

6 Depalma MJ, Ketchum JM, Trussell BS, Saullo TR, Slipman CW. Does the location of low back pain predict its source? Pm & R 2011;3(1):33-39.

7 Donelson R, Aprill C, Medcalf R, Grant W. A prospective study of centralization of lumbar and referred pain. A predictor of symptomatic discs and anular competence. Spine (Phila Pa 1976) 1997;22(10):1115-1122.

8 Dreyfuss P, Michaelsen M, Pauza K, McLarty J, Bogduk N. The value of medical history and physical examination in diagnosing sacroiliac joint pain. Spine (Phila Pa 1976) 1996;21(22):2594-2602.

9 Freiermuth D, Kretzschmar M, Bilecen D, Schaeren S, Jacob AL, Aeschbach A, Ruppen W. Correlation of (99m) Tc-DPD SPECT/CT Scan Findings and Diagnostic Blockades of Lumbar Medial Branches in Patients with Unspecific Low Back Pain in a Randomized-Controlled Trial. Pain Medicine 2015;16(10):1916-1922.

10 Hanna H, Tommy H, Hebelka H, Hansson T. HIZ's relation to axial load and low back pain: investigated with axial loaded MRI and pressure controlled discography. European Spine Journal 2013;22(4):734-739.

11 Holder LE, Machin JL, Asdourian PL, Links JM, Sexton CC. Planar and high-resolution SPECT bone imaging in the diagnosis of facet syndrome. J Nucl Med 1995;36(1):37-44.

12 Ito M, Incorvaia KM, Yu SF, Fredrickson BE, Yuan HA, Rosenbaum AE. Predictive signs of discogenic lumbar pain on magnetic resonance imaging with discography correlation. Spine (Phila Pa 1976) 1998;23(11):1252-1258; discussion 1259-1260.

13 Lam KS, Carlin D, Mulholland RC. Lumbar disc high-intensity zone: the value and significance of provocative discography in the determination of the discogenic pain source. Eur Spine J 2000;9(1):36-41.

14 Laslett M, Aprill CN, McDonald B, Oberg B. Clinical predictors of lumbar provocation discography: a study of clinical predictors of lumbar provocation discography. Eur Spine J 2006;15(10):1473-1484.

15 Laslett M, Aprill CN, McDonald B, Young SB. Diagnosis of sacroiliac joint pain: validity of individual provocation tests and composites of tests. Man Ther 2005;10(3):207-218.

16 Laslett M, Oberg B, Aprill CN, McDonald B. Centralization as a predictor of provocation discography results in chronic low back pain, and the influence of disability and distress on diagnostic power. Spine J 2005;5(4):370-380.

17 Laslett M, Young SB, Aprill CN, McDonald B. Diagnosing painful sacroiliac joints: A validity study of a McKenzie evaluation and sacroiliac provocation tests. Aust J Physiother 2003;49(2):89-97.

18 Lei D, Rege A, Koti M, Smith FW, Wardlaw D. Painful disc lesion: can modern biplanar magnetic resonance imaging replace discography? Journal of Spinal Disorders & Techniques 2008;21(6):430-435.
19 Lim CH, Jee WH, Son BC, Kim DH, Ha KY, Park CK. Discogenic lumbar pain: association with MR imaging and CT discography. Eur J Radiol 2005;54(3):431-437.
20 Maigne JY, Boulahdour H, Chatellier G. Value of quantitative radionuclide bone scanning in the diagnosis of sacroiliac joint syndrome in 32 patients with low back pain. Eur Spine J 1998;7(4):328-331.
21 Mekhail N, Saweris Y, Sue Mehanny D, Makarova N, Guirguis M, Costandi S. Diagnosis of Sacroiliac Joint Pain: Predictive Value of Three Diagnostic Clinical Tests. Pain Pract 2021;21(2):204-214.
22 Nejati P, Sartaj E, Imani F, Moeineddin R, Nejati L, Safavi M. Accuracy of the Diagnostic Tests of Sacroiliac Joint Dysfunction. J Chiropr Med 2020;19(1):28-37.
23 O'Neill C, Kurgansky M, Kaiser J, Lau W. Accuracy of MRI for diagnosis of discogenic pain. Pain Physician 2008;11(3):311-326.
24 Parker LM, Murrell SE, Boden SD, Horton WC. The outcome of posterolateral fusion in highly selected patients with discogenic low back pain. Spine (Phila Pa 1976) 1996;21(16):1909-1916; discussion 1916-1907.
25 Ricketson R, Simmons JW, Hauser BO. The prolapsed intervertebral disc. The high-intensity zone with discography correlation. Spine (Phila Pa 1976) 1996;21(23):2758-2762.
26 Saifuddin A, Braithwaite I, White J, Taylor BA, Renton P. The value of lumbar spine magnetic resonance imaging in the demonstration of anular tears. Spine (Phila Pa 1976) 1998;23(4):453-457.
27 Schellhas KP, Pollei SR, Gundry CR, Heithoff KB. Lumbar disc high-intensity zone. Correlation of magnetic resonance imaging and discography. Spine (Phila Pa 1976) 1996;21(1):79-86.
28 Schneider BJ, Ehsanian R, Rosati R, Huynh L, Levin J, Kennedy DJ. Validity of Physical Exam Maneuvers in the Diagnosis of Sacroiliac Joint Pathology. Pain Med 2020;21(2):255-260.
29 Slipman CW, Sterenfeld EB, Chou LH, Herzog R, Vresilovic E. The value of radionuclide imaging in the diagnosis of sacroiliac joint syndrome. Spine (Phila Pa 1976) 1996;21(19):2251-2254.
30 Smith BM, Hurwitz EL, Solsberg D, Rubinstein D, Corenman DS, Dwyer AP, Kleiner J. Interobserver reliability of detecting lumbar intervertebral disc high-intensity zone on magnetic resonance imaging and association of high-intensity zone with pain and anular disruption. Spine (Phila Pa 1976) 1998;23(19):2074-2080.
31 Stanford G, Burnham RS. Is it useful to repeat sacroiliac joint provocative tests post-block? Pain Medicine 2010;11(12):1774-1776.
32 van der Wurff P, Buijs EJ, Groen GJ. A multitest regimen of pain provocation tests as an aid to reduce unnecessary minimally invasive sacroiliac joint procedures. Arch Phys Med Rehabil 2006;87(1):10-14.
33 Vanharanta H, Ohnmeiss DD, Aprill CN. Vibration pain provocation can improve the specificity of MRI in the diagnosis of symptomatic lumbar disc rupture. Clin J Pain 1998;14(3):239-247.
34 Weishaupt D, Zanetti M, Hodler J, Min K, Fuchs B, Pfirrmann CW, Boos N. Painful Lumbar Disk Derangement: Relevance of Endplate Abnormalities at MR Imaging. Radiology 2001;218(2):420-427.
35 Yoshida H, Fujiwara A, Tamai K, Kobayashi N, Saiki K, Saotome K. Diagnosis of symptomatic disc by magnetic resonance imaging: T2-weighted and gadolinium-DTPA-enhanced T1-weighted magnetic resonance imaging. J Spinal Disord Tech 2002;15(3):193-198.
36 Young S, Aprill C, Laslett M. Correlation of clinical examination characteristics with three sources of chronic low back pain. Spine J 2003;3(6):460-465.

Chapter Six:

Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review

Statement from co-authors confirming authorship contribution of the PhD candidate

The co-authors of the paper *"Han CS, Maher CG, Steffens D, Diwan A, Magnussen J, Hancock EC, Hancock MJ. Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review. J Physiother. 2023 Apr;69(2):79-92. doi: 10.1016/j.jphys.2023.02.007. Epub 2023 Mar 11. PMID: 36914521."* confirm that Christopher S Han has provided the following contributions to the study:

- Conception and design of the research
- Data acquisition
- Data analysis and interpretation of findings
- Writing of the manuscript and critical appraisal of the content

Christopher S Han _____ Date: 22/11/2023

Professor Chris G Maher _____ Date: 22/11/2023

Professor Mark J Hancock _____ Date: 22/11/2023

Journal of Physiotherapy ■ (2023) ■–■

AUSTRALIAN PHYSIOTHERAPY ASSOCIATION

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Research

Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review

Christopher S Han [a], Christopher G Maher [a], Daniel Steffens [b,c], Ashish Diwan [d], John Magnussen [e], Emma C Hancock [f], Mark J Hancock [g]

[a] *Institute for Musculoskeletal Health, The University of Sydney and Sydney Local Health District, Sydney, Australia;* [b] *Surgical Outcomes Research Centre (SOuRCe), Royal Prince Alfred Hospital, Sydney, Australia;* [c] *Central Clinical School, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia;* [d] *Spine Labs, St. George and Sutherland Clinical School, University of New South Wales, Sydney, Australia;* [e] *Department of Clinical Medicine, Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia;* [f] *University of New South Wales, Sydney, Australia;* [g] *Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia*

KEY WORDS

Magnetic resonance imaging
Predict
Diagnosis
Low back pain
Musculoskeletal

ABSTRACT

Questions: Do magnetic resonance imaging (MRI) findings predict future low back pain (LBP), associated disability and global recovery in people with current LBP? Do MRI findings predict these outcomes in people with no current LBP? Do MRI findings predict these outcomes in a mixed sample of people with and without current LBP? **Design**: This review is an update of a previous systematic review investigating the relationship between lumbar spine MRI findings and future LBP. **Participants**: People with or without LBP with lumbar MRI scans. **Outcome measures**: MRI findings, pain and disability. **Results**: Of the included studies, 28 reported on participants with current LBP, eight reported on participants with no LBP and four reported on a mixed sample. Most results were based on single studies and did not demonstrate clear relationships between MRI findings and future LBP. In populations with current LBP, pooling demonstrated that the presence of Modic type 1 changes alone or Modic type 1 and 2 changes were associated with slightly worse pain or disability outcomes in the short term, and the presence of disc degeneration was associated with worse pain and disability outcomes in the long term. In populations with current LBP, pooling demonstrated no evidence of an association between the presence of nerve root compression and disability outcomes in the short term, and no evidence of an association between the presence of disc height reduction, disc herniation, spinal stenosis, high-intensity zone and clinical outcomes in the long term. In populations with no LBP, pooling demonstrated that the presence of disc degeneration may increase the likelihood of experiencing pain in the long term. In mixed populations, no pooling was possible; however, single studies demonstrated that Modic type 1, 2 or 3 changes and disc herniation were each associated with worse pain in the long term. **Conclusions**: The results suggest that some MRI findings may have weak associations with future LBP; however, larger high-quality studies are needed to resolve uncertainty. **Systematic review registration**: PROSPERO CRD42021252919. **[Han CS, Maher CG, Steffens D, Diwan A, Magnussen J, Hancock EC, Hancock MJ (2023) Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review. *Journal of Physiotherapy* ■:■–■]**

Introduction

Low back pain (LBP) continues to be the leading cause of years lived with disability, according to The Global Burden of Disease Study.[1] After excluding specific causes of LBP (eg, radiculopathy) and serious pathologies (eg, lumbar spine fracture), the majority (90 to 95%) of people experiencing LBP are classified as having non-specific LBP.[2] This label acknowledges the inability to identify the specific cause of most LBP.[2] One view is that if the nociceptive source of pain could be identified, more specific and logical treatments could be developed and provided. A potential group of features that could be used with this approach are structural changes observed on lumbar spine imaging. A better understanding of magnetic resonance imaging (MRI) findings and their relevance to LBP could potentially be used to define subgroups that follow a different natural course and to moderate the effects of specific treatments.

The importance of various MRI findings and their link to the aetiology of LBP has been challenged, as structural changes on imaging can be present in both symptomatic and asymptomatic people.[3] For example, one review (14 cross-sectional studies, n = 3,097) found that disc degeneration on imaging was present in 57% of symptomatic people, but also present in 34% of asymptomatic people.[4] However, it is important to note that the prevalence of disc degeneration was higher in people with a history of LBP (compared with those with no history of LBP), as was the prevalence of disc bulge, disc extrusion, disc protrusion, Modic type 1 changes and spondylolysis.[4] Further

https://doi.org/10.1016/j.jphys.2023.02.007

complicating interpretation of MRI findings is that the prevalence of most imaging findings increases substantially with age.[4]

A limitation of most previous research studies investigating the relationship between structural changes on MRI and LBP is that they use cross-sectional[5,6] designs, which only provide weak evidence of a causal relationship.[6] Most imaging studies have also been limited by small samples, and inconsistent measures of MRI findings and LBP status. Acknowledging the limitations of cross-sectional designs, other research studies have used longitudinal designs to investigate the relationship between MRI findings and future LBP.

Our previous systematic review investigating the relationship between MRI findings and future LBP found that definitive conclusions were not possible due to limitations such as a paucity of high-quality studies and significant heterogeneity between studies.[6] However, single studies in our previous review did report significant associations between Modic type 1 changes and future pain and between disc degeneration and future disability, in samples with current LBP.[6] Very few studies in our previous review investigated whether MRI findings predicted LBP in populations that were pain free at baseline.[6] Given that our review is now 8 years old, an update of the review with recently published longitudinal studies was performed.

Therefore, the specific research questions for this systematic review were:

1. Do MRI findings predict future LBP, associated disability and global recovery in people with current LBP?
2. Do MRI findings predict these outcomes in people with no current LBP?
3. Do MRI findings predict these outcomes in a mixed sample of people with and without current LBP?

Methods

A review protocol was specified in advance and registered on PROSPERO (CRD42021252919).

Search strategy

MEDLINE, CINAHL and EMBASE were searched for the period January 2012 to 24 November 2021 with the search strategy used in our earlier review.[6] The complete search strategies from all databases are presented in Appendix 1 on the eAddenda. The reference lists of identified papers were examined and forward citation searching was also performed.

Inclusion criteria

Studies were required to meet all the following criteria to be included:

Longitudinal cohort study: This included secondary analyses of randomised controlled trials. Trials where the intervention provided was designed to alter lumbar spine MRI findings (eg, surgery or antibiotic treatment for Modic changes) were excluded, unless the data from the other treatment groups could be extracted separately. We reviewed all studies excluded by the previous study authors at the full-text stage from the review by Steffens et al[6] to include any studies previously excluded for being retrospective studies.

MRI undertaken: Participants must have undergone a baseline MRI and had at least one lumbar spine finding reported; for example, but not limited to, disc degeneration, disc herniation, disc protrusion, facet joint arthropathy, Modic changes or high-intensity zone.

Outcome measure relevant to LBP reported: Studies must have reported a clinically relevant outcome measure at follow-up, such as pain, disability and a global measure of recovery status. A follow-up time point of $\geq$ 4 weeks was accepted for both symptomatic and asymptomatic population groups.

Given that the review by Steffens et al[6] used a follow-up time point of $\geq$ 3 months, all studies excluded at the full-text stage were searched to include any studies previously excluded for having a follow-up time point $<$ 3 months. It is unlikely that the review by Steffens et al[6] would have excluded studies during the title and abstract screening stage when assessing study eligibility based on minimum follow-up time points.

When studies reported multiple measures from the domain of pain outcomes, the Numeric Pain Rating Scale or Visual Analogue Scale was preferentially used.[7] For disability, any version of the Oswestry Disability Index was preferentially used if available. If the Oswestry Disability Index was not reported, data were extracted from the Roland Morris Disability Questionnaire.[7] If neither were used, whichever disability outcome measure was reported was extracted. For global measure of recovery status, whichever outcome measure was reported was extracted. Outcome data were re-scaled to a common 0-to-100 scale to facilitate between-study comparisons.

Association reported: The association between MRI findings and LBP outcomes must have been reported. If associations were not reported, sufficient raw data to calculate a measure of association must have been reported. Reporting just a *p*-value was insufficient to meet this criterion.

Exclusion criteria

Studies including participants with serious or specific pathologies such as tumours, fractures, inflammatory arthritis and cauda equina were excluded. One reviewer (CSH) screened all titles to exclude irrelevant studies. Two reviewers (CSH and ECH) independently screened abstracts to exclude irrelevant studies. Two reviewers (CSH and ECH) then independently reviewed full texts for eligibility based on the inclusion criteria. Any discrepancies were resolved through discussion and a third reviewer (CGM or MJH). Relevant authors were contacted via email if any additional information was required.

Data extraction

All relevant data were extracted independently by two reviewers. The data were categorised into two groups based upon whether study participants had pain at baseline. In the mixed sample, if a clear majority ($\geq$ 95%) of participants had either pain or no pain at baseline, the data were categorised based on the pain status of the majority. Data extracted from the included studies were number of participants, participant age, presence of pain at baseline, follow-up duration, MRI findings, outcome measures, and strength of associations between MRI findings and outcome measures. When sufficient raw data were available, odds ratios and 95% confidence intervals were calculated.

Methodological quality appraisal

We originally intended to assess the methodological quality of each study using the QUality In Prognostic Studies tool; however, the Altman[8] checklist was used as it was simpler to use and also used in our previous review. The methodological quality of each study was assessed using the standardised checklist of pre-defined criteria[9,10] used in the review by Steffens et al.[6] This checklist is a modified version based on theoretical considerations and methodological aspects described by Altman et al.[8] These criteria are presented in Box 1.

Two reviewers independently assessed each included study. Each criterion was scored positive 'yes' or negative 'no'. A positive score indicated sufficient information and a positive assessment. Any disagreements in scores were resolved through discussion or a third reviewer (CGM or MJH). No overall methodological quality score was used as an inclusion criterion.

Box 1. Criteria used for appraisal of methodological quality.

- definition of study sample (description of participant source and inclusion and exclusion criteria)
- representative sample of the target sample (participants selected by random selection or as consecutive cases)
- follow-up rate ≥ 80% (outcome data were available for ≥ 80% of participants at 3-month follow-up or later)
- adequate follow-up time (≥ one prognostic outcome was followed up at 3 months or later)
- interpretable prognostic outcomes available (raw data, percentages, survival rates or continuous outcome reported)
- blinding (assessor unaware of ≥ one prognostic factor used to predict prognostic outcome, at the time prognostic outcome was measured)

Data analysis

Following data extraction, study level associations were collated. Data were pooled using meta-analysis with a random effects model when possible. Commercial software[a] was used for meta-analysis. Due to the heterogeneity between studies (ie, MRI findings, clinical outcome measures, statistical measure used to assess association of MRI finding), some results were also presented descriptively. Study results were reported and analysed separately based on the presence or absence of pain at baseline.

All positive scores for continuous measures (eg, mean difference) indicated that the exposure group had a worse outcome at follow-up compared with the comparison group. Odds ratios or hazard ratios > 1 indicated greater likelihood of poorer outcome in the exposure group compared with the comparison group. Short-term outcomes were defined as a follow-up time point < 12 months, and the closest time point to 3 months was taken. Long-term outcomes were defined as a follow-up time point ≥ 12 months, and the closest time point to 12 months was taken.

Associations were defined as: 'no evidence of an association' if the effect sizes were small (MD < 10 points on a 100-point scale, OR/HR < 1.5) and confidence intervals spanned no association (or where confidence intervals were not reported, the result was not statistically significant). If the effect sizes were small but the confidence intervals did not span no association (or the result was statistically significant), this was defined as a 'weak association' and the term 'slightly' was used to imply a small effect. If the effect size was moderate to large (MD >10 points on a 100-point scale, OR/HR > 1.5) and confidence intervals did not span no association (or the result was statistically significant), this was defined as having an 'association'. If the effect size was moderate to large but confidence intervals spanned no association (or the result was not statistically significant), the term 'may' was used to imply a potential association.

Results

Selection of studies

After removing duplicates (n = 976) the search identified 9,949 citations. Of these, 6,919 were deemed clearly irrelevant based on screening by title and a further 2,882 studies were excluded based on title and abstract. Two studies[11,12] originally excluded in the Steffens et al review[6] were included for full-text screening as they potentially met the criteria. Following full-text review by two independent authors, 27 new studies[13–39] and the 12 studies from the previous review[5,40–50] were included for analysis (Figure 1). One study presented data separately for populations with current LBP and populations with no LBP;[25] therefore, this study was separated into two strata for analysis: Kasch 2022a and Kasch 2022b.[25] A total of 39 studies were included for analysis. Excluded studies and the primary reasons for exclusion are presented in Appendix 2 on the eAddenda.

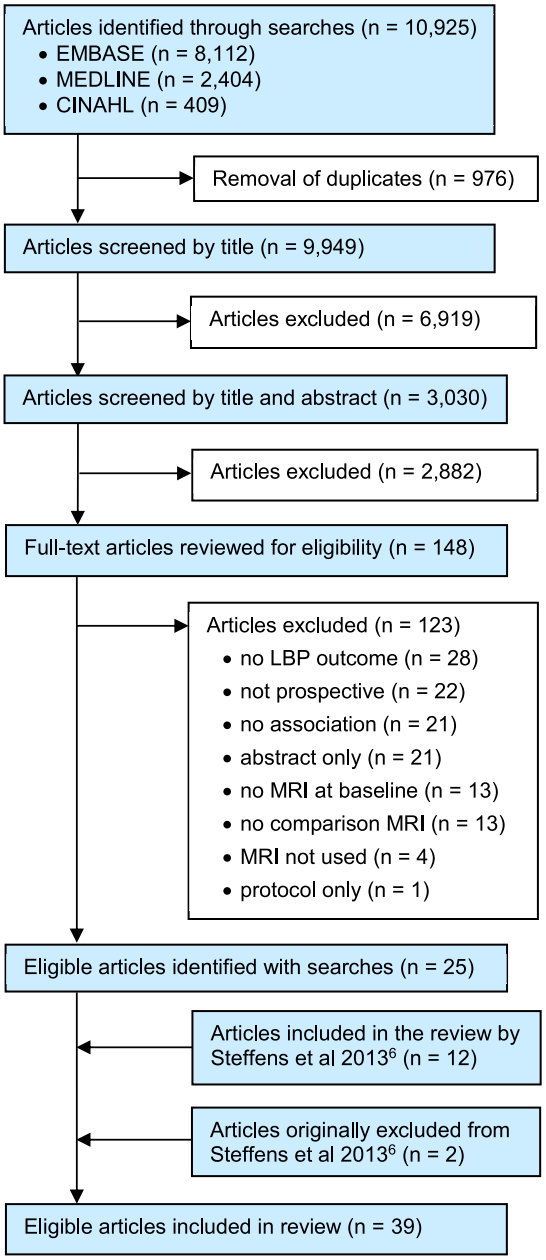


Figure 1. Flow of studies through the review.
LBP = low back pain, MRI = magnetic resonance imaging.

Management of missing information

Five authors of conference abstracts[51–55] or protocols[56] were contacted for full-text articles, and two full texts were received.[52,57] Six authors[14,17,21,26,47,58] were contacted regarding missing data and/or discrepancies; three[14,17,21] were able to provide additional data.

Methodological quality

The results of the methodological quality assessment for all 39 studies are presented in Table 1. Thirty-six studies provided an adequate description of the study sample. Nineteen studies reported selecting participants by random selection or as consecutive cases. Twenty-six studies reported > 80% follow-up. All studies reported adequate follow-up time. All studies reported a prognostic outcome at ≥ 3 months. Twenty-nine studies reported assessor blinding.

Table 1
Methodological quality assessment of included studies.

Author, year	Definition of study sample[a]	Representative sample[b]	Follow-up rate ≥ 80%[c]	Adequate follow-up time[d]	Interpretable prognostic outcomes available[e]	Blinding[f]
Boos, 2000[40]	Yes	No	Yes	Yes	Yes	Yes
Borenstein, 200[15]	Yes	No	No	Yes	Yes	Yes
Elfering, 2002[43]	No	No	Yes	Yes	Yes	No
Carragee, 2005[42]	Yes	Yes	Yes	Yes	Yes	Yes
Jarvik, 2005[45]	Yes	Yes	Yes	Yes	Yes	No
Modic, 2005[50]	Yes	No	Yes	Yes	Yes	Yes
Carragee, 2006[41]	Yes	Yes	Yes	Yes	Yes	Yes
Kleinstuck, 2006[48]	Yes	No	Yes	Yes	Yes	Yes
McNee, 2011[49]	Yes	Yes	No	Yes	Yes	Yes
Hellum, 2012[44]	Yes	No	Yes	Yes	Yes	Yes
Jensen, 2012[46]	Yes	No	Yes	Yes	Yes	Yes
Keller, 2012[47]	Yes	No	No	Yes	Yes	Yes
Wilkens, 2013[39]	Yes	Yes	Yes	Yes	Yes	No
el Barzouhi, 2014[18]	Yes	No	Yes	Yes	Yes	No
Jensen, 2014[23]	Yes	No	Yes	Yes	Yes	Yes
Peterson, 2014[32]	Yes	Yes	No	Yes	Yes	Yes
Schistad, 2014[35]	Yes	No	Yes	Yes	Yes	Yes
Hancock, 2015[20]	Yes	Yes	Yes	Yes	Yes	Yes
Jarvinen, 2015[22]	Yes	Yes	Yes	Yes	Yes	Yes
Jensen, 2015[24]	Yes	Yes	Yes	Yes	Yes	Yes
Kerr, 2015[26]	Yes	No	No	Yes	Yes	No
Määttä, 2015[31]	Yes	No	No	Yes	Yes	Yes
Annen, 2016[14]	Yes	Yes	Yes	Yes	Yes	Yes
de Schepper, 2016[17]	Yes	Yes	Yes	Yes	Yes	Yes
Suri, 2016[36]	Yes	Yes	Yes	Yes	Yes	No
Hancock, 2017[19]	Yes	Yes	Yes	Yes	Yes	Yes
Koivisto, 2017[28]	Yes	No	Yes	Yes	Yes	Yes
Romero-Munoz, 2017[33]	No	Yes	Yes	Yes	Yes	Yes
Tonosu, 2017[37]	Yes	No	No	Yes	Yes	Yes
Cai, 2018[15]	Yes	No	Yes	Yes	Yes	Yes
Annen, 2018[13]	Yes	Yes	No	Yes	Yes	Yes
Konstantinou, 2018[29]	Yes	Yes	No	Yes	Yes	No
Chen, 2019[16]	Yes	No	Yes	Yes	Yes	No
Kesikburun, 2019[27]	Yes	Yes	No	Yes	Yes	Yes
Harrisson, 2019[54]	Yes	No	No	Yes	Yes	No
Kuligowski, 2020[30]	Yes	No	Yes	Yes	Yes	No
Sääksjärvi, 2020[34]	Yes	No	No	Yes	Yes	Yes
Kasch, 2022[25]	No	Yes	No	Yes	Yes	Yes
Udby, 2021[38]	Yes	Yes	Yes	Yes	Yes	Yes

[a] Description of participant source and inclusion and exclusion criteria.
[b] Participants selected by random selection or as consecutive cases.
[c] Outcome data were available for ≥ 80% of participants at 3-month follow-up or later.
[d] At least one prognostic outcome was followed up at 3 months or later.
[e] Raw data, percentages, survival rates or continuous outcome reported.
[f] Assessor of MRI unaware of ≥ one prognostic factor, used to predict prognostic outcome, at the time prognostic outcome was measured.

Study characteristics

A full description of included studies is presented in Table 2. Twenty-eight studies[13–18,21–30,32,33,35,36,38,39,44,46–50] included participants with current LBP, eight studies[5,19,20,25,37,40,43,45] included participants with no LBP and four studies[31,34,41,42] enrolled a mixed participant sample. The mean participant age ranged from 29 to 62 years. The follow-up period ranged from 4 weeks to 30 years.

Association of MRI findings and clinical outcomes in people with current LBP

Twenty-eight studies[13–18,21–30,32,33,35,36,38,39,44,46–50] reported a total of 377 associations between MRI findings and clinical outcomes in participants with current LBP (Appendix 3 on the eAddenda). The MRI findings presented in three or more studies or where associations could be pooled across at least two studies are presented below. Results of associations between MRI findings and pain or disability outcomes are presented in Tables 3 and 4, respectively. All MRI findings and outcomes are presented in Appendix 3.

Association between Modic changes and pain and disability

Nineteen studies[13–16,18,22–25,28,32,33,35,38,39,44,46–48] assessed the associations between Modic changes (type 1, type 2, type 3, or type 1 and/or 2) and pain[13–16,18,22–25,28,32,33,35,46–48] or disability[14,16,22,23,28,33,38,39,44,47,48] (Tables 3 and 4). Results were able to be pooled for 19 different comparisons (which included 55 associations from individual studies) in total (Table 5). Only results of associations investigating different types of Modic changes compared with no Modic changes for pain or disability outcomes are presented below. Results of all pooled associations are presented in Table 5.

Any Modic change: Pooling demonstrated those with Modic changes (type 1 and 2, type 1, and type 2) compared with no Modic change had slightly worse pain (MD 6.6, 95% CI 4.3 to 8.9)[13–16,32,35] and disability outcomes (MD 4.6, 95% CI 0.3 to 8.9)[14,16] in the short term. In the long term, pooling demonstrated no evidence of an association between Modic changes (Modic type 1 and Modic type 2) compared with no Modic changes for pain (MD 0.0, 95% CI −6.5 to 6.4)[14,23–25,35,47] and disability outcomes (MD −0.3, 95% CI −3.7 to 3.2)[14,23,39,47] (Table 5).

Modic type 1: Pooling demonstrated that those with Modic type 1 changes compared with no Modic changes had slightly worse pain outcomes (MD 4.4, 95% CI 1.1 to 7.6);[13–16,35] however, there was no evidence of an association for disability outcomes (MD 3.1, 95% CI −4.3 to 10.5)[14,16] in the short term. In the long term, pooling demonstrated no evidence of an association between Modic type 1 changes compared with no Modic changes and pain (MD 1.2, 95% CI −9.1 to 11.4)[14,23–25,35,47] and disability outcomes (MD −2.4, 95% CI −8.8 to 4.0).[14,23,39,47] Pooling also demonstrated no evidence of an association between Modic type 1 changes compared with no Modic changes and the presence of pain (OR 1.1, 95% CI 0.5 to 2.4)[18,46] in the long term.

Table 2
Individual study characteristics.

Study	Participants		Follow-up duration	Clinical outcome	MRI finding
	N (% female)	Age (y), mean (SD)[a]			
Populations with current LBP					
Modic, 2005[50]	246 (58%)	43 (NR)	2 y	• Disability (RMDQ)	Canal stenosis, nerve root compression, disc herniation
Kleinstuck, 2006[48]	53 (49%)	44 (not reported)	1 y	• Pain last 2 wk (VAS) • Disability (RMDQ)	High-intensity zone, disc bulge, disc degeneration, Modic type 1, Modic type 2
McNee, 2011[49]	240 (54%)	NR (range 20 to 60)	22 mo (mean)	• Pain > 14 d in past 4 wk • Disability past 4 wk (RMDQ)	Number of MRI abnormalities, disc degeneration
Hellum, 2012[44]	66 (58%)	42 (NR)	1 y	• Disability (ODI)	Modic type 1, Modic type 2, disc degeneration (height reduction), disc degeneration (signal intensity), facet joint arthropathy, high-intensity zone
Jensen, 2012[46]	96 (69%)	46 (range 21 to 60)	14 mo	• Pain (NPRS)	Modic type 1, Modic type 1 (change in size)
Keller, 2012[47]	269 (50%)	50 (range 20 to 60)	1 y	• Patient global impression of improvement (PGI-I)	Modic type 1, Modic type 2
Wilkens, 2013[39]	250 (48%)	49 (11)	1 y	• Disability (RMDQ)	High-intensity zone, Modic type 1, Modic type 2, and Modic type 1, 2 or 3
el Barzouhi, 2014[18]	95 (29%)	44 (10)	4 wk 3 mo 6 mo 1 y	• Disabling LBP ≥ 40/100 mm • Pain (VAS)	Vertebral endplate signal changes type 1, vertebral endplate signal changes type 2, vertebral endplate signal changes type 1 and 2
Jensen, 2014[23]	141 (53%)	42 (11)	1 y	• Pain (LBP rating scale)	Disc nucleus signal, lumbar osteophytes, disc height reduction, high-intensity zone, disc herniation, nerve root compression, spina stenosis, Modic type 1, Modic type 2, and Modic type 1 and 2
Peterson, 2014[32]	345 (50%)	60 (NR)	1 mo	• Patient impression of change (PGIC)	Modic type 1 and 2
Schistad, 2014[35]	235 (47%)	41 (11)	6 wk 6 mo 1 y	• Pain (VAS) • Pain experience (McGill sensory pain score)	Modic type 1, Modic type 2 and 3
Jarvinen, 2015[22]	64 (87%)	44 (10)	2 y	• Pain (NRS) • (Disability) ODI	Change in Modic type 1 extent, change in Modic type 2 extent
Jensen, 2015[24]	96 (69%)	46 (range 21 to 60)	14 mo	• Change in pain (NRS)	Modic type 1, large Modic change (≥ 25% vertebral height), large Modic type 1, severe disc degeneration (≥ 3 Pfirrmann), large disc herniation
Kerr, 2015[26]	392 (41%)	44 (12)	8 y	• Disability (ODI)	Posterolateral herniation, other herniation location, extrusion, protrusion, sequestration
Annen, 2016[14]	72 (23%)	42 (11)	1 mo 3 mo 6 mo 1 y	• Improvement (PGIC) • Back pain (NRS) • Leg pain (NRS) • Disability (ODI)	Modic type 1, Modic type 2, Modic type 1 and 2
de Schepper, 2016[17]	683 (47%)	50 (13)	1 y	• Recovery (strongly improved/ completely improved on GPE)	Disc bulging, disc herniation, nerve root compression, spinal stenosis, spondylolisthesis
Suri, 2016[36]	341 (62%)	44 (12)	4 y	• Back pain recurrence (LBP bothersome) • Leg pain recurrence (Sciatic bothersome index)	Extrusion, sequestration, posterolateral disc herniation
Koivisto, 2017[28]	20 (45%)	52 (NR)	51 wk	• Change in LBP intensity (VAS) • Disability (ODI)	Change in Modic type 1 extent, change in Modic type 2 extent
Romero-Munoz, 2017[33]	70 (66%)	57 (range 43 to 67)	10 y	• Back pain (VAS) • Leg pain (VAS) • Disability (ODI)	Modic type 1, Modic type 2, Modic type 3
Annen, 2018[13]	112 (42%)	45 (16)	1 mo 3 mo	• Pain (NRS) • Bournemouth Questionnaire • Improvement (PGIC)	Modic type 1, Modic type 2, and Modic type 1 and 2
Cai, 2018[15]	37 (39%)	62 (11)	6 mo	• Pain (VAS)	Mild disc degeneration, severe disc degeneration, Modic type 1

Table 2 (*Continued*)

Study	Participants		Follow-up duration	Clinical outcome	MRI finding
	N (% female)	Age (y), mean (SD)[a]			
Konstantinou, 2018[29]	609 (63%)	50 (14)	4 mo 12 mo	• Disability (RMDQ)	Nerve root compression
Chen, 2019[16]	129 (57%)	42 (range 21 to 67)	3 mo 6 mo	• Pain (VAS) • Disability (ODI)	Modic type 1, Modic type 2
Kesikburun, 2019[27]	40 (45%)	50 (14)	17 mo (SD 7)	• Pain (NRS) • Disability (ODI)	Disc herniation
Harrisson, 2019[54]	402 (63%)	50 (14)	4 mo 1 y 3 y	• Mean pain intensity (0 to 10)	Nerve root compression
Kuligowski, 2020[30]	30 (60%)	29 (3)	4 wk	• Pain (NRS) • Disability (ODI)	Protrusion, extrusion
Kasch, 2022[a]	1,997 (54%)	53 (14)	6 y	• Pain (NRS)	Disc degeneration, disc height loss, disc herniation, high-intensity zone, spondylolisthesis, Modic type 1, Modic type 2 or 3, hypertrophy ligamentus flava, Schmorl lesion, spinal canal stenosis, 1 MRI finding, 2 MRI findings, 3 MRI findings, 4 MRI findings, ≥ 5 MRI findings
Udby, 2021[38]	170 (54%)	53 (range 31 to 70)	13 y	• Disability (RMDQ)	Modic change (1, 2, 3), disc degeneration, facet joint degeneration
Populations with no LBP					
Boos, 2000[40]	46 (NR)	NR (range 20 to 50)	62 mo (mean)	• Pain (NQ)	Disc herniation, nerve root contact, disc degeneration
Borenstein, 200[15]	67 (NR)	42 (NR)	7 y	• Pain (ordinal scale)	MRI abnormalities (change)
Elfering, 2002[43]	41 (27%)	NR (range 20 to 50)	62 mo (mean)	• Pain (NQ)	Disc degeneration (change)
Jarvik, 2005[45]	148 (11%)	54 (range 36 to 71)	3 y	• Pain (PFI)	Disc herniation, nerve root contact, canal stenosis
Hancock, 2015[20]	76 (39%)	46 (13)	1 y	• LBP recurrence (return of LBP lasting at least 24 h, pain intensity ≥ 3/10 NRS) • Recurrence of activity limited LBP	Disc degeneration ≥ 3, high-intensity zone, Modic change, disc herniation, facet joint arthrosis, spondylolisthesis
Hancock, 2017[19]	76 (39%)	46 (13)	52 wk	• Days of LBP recurrence (adaptation of item 8 of SF-36)	Pfirrmann DD score ≥ 3, disc height loss, disc annular tear, high-intensity zone, disc herniation, spondylolisthesis, facet joint osteoarthritis, Modic change, and 1, 2 or 3, aggregate MRI findings
Tonosu, 2017[37]	49 (51%)	45 (9)	10 y	• Back pain in the past 1 mo	Disc degeneration, disc bulge, high-intensity zone
Kasch, 2022[b]	1,366 (46%)	53 (14)	6 y	• Pain (NRS)	Disc degeneration, disc height loss, disc herniation, high-intensity zone, spondylolisthesis, Modic type 1, Modic type 2 or 3, hypertrophy ligamentus flava, Schmorl lesion, spinal canal stenosis, 1 MRI finding, 2 MRI findings, 3 MRI findings, 4 MRI findings, ≥ 5 MRI findings
Populations with a mixed sample					
Carragee, 2005[42]	100 (37%)	42 (NR)	63 mo (mean)	• Pain (remission)	Disc herniation
Carragee, 2006[41]	200 (40%)	39 (NR)	5 y	• Pain (NRPS) • Disability (ODI)	Disc degeneration, endplate changes, canal stenosis
Määttä, 2015[31]	429 (98%)	54 (NR)	10 y	• Disabling low back pain lasting > 1 mo	DD summary score (0 to 60), Schmorl node, Modic change
Sääksjärvi 2020,[34]	35 (0%)	51 (1)	30 y	• Pain (VAS) • Disability (ODI)	Single intensity decrease > 60%, Modic change, disc protrusion or bulge, severe body endplate lesion, disc height reduction, annular fissure, high-intensity zone

GPE = global perceived effect, LBP = low back pain, MRI = magnetic resonance imaging, NPRS = numeric pain rating scale, NQ = Nordic questionnaire, NRS = numeric rating scale, ODI = Oswestry Disability Index, PFI = Pain Frequency Index, RMDQ = Roland Morris Disability Questionnaire, SF-36 = 36 Item Short Form Survey, VAS = visual analogue score.
[a] Unless otherwise stated.

Fourteen studies[14,18,22–25,28,33,35,38,39,44,47,48] investigated the remaining 46 associations between Modic change (any type) compared with no Modic changes for pain or disability outcomes in populations with LBP. Pooling was not possible due to heterogeneity in outcome measures or measures of association (Tables 3 and 4). Point estimates in 40 of the 46 associations demonstrated no evidence of an association between Modic changes (type 1 and 2, type 1 alone, type 2 alone, or type 3 alone) compared with no Modic

Table 3
Association between MRI findings and clinical outcomes in populations with current low back pain in the short term.

Exposure	Comparison	Clinical outcome (follow up)[a]	Difference in means (95% CI)[b]
Disc bulge[48]	No disc bulge	Back pain (3 mo)[c]	0.1 (p = 0.6)[d]
Disc bulge[48]	No disc bulge	Disability (6 wk)[c]	0.2 (p = 0.1)[d]
Disc degeneration[48]	No disc degeneration	Back pain (3 mo)[c]	0.1 (p = 0.6)[d]
Disc degeneration (severe)[15]	Disc degeneration (mild)	Back pain (6 mo)	−8.5 (−23.4 to 6.4)[e]
Disc degeneration[48]	No disc degeneration	Disability (3 mo)[c]	0.1 (p = 0.7)[d]
Disc herniation[50]	No disc herniation	Disability (6 wk)	2.7 (1.4 to 5.1) (OR)
Disc protrusion[30]	Disc extrusion	Back pain (1 mo)	30.0 (NR)[f]
Disc protrusion[30]	Disc extrusion	Disability (1 mo)	18.0 (NR)[f]
High intensity zone[48]	No high intensity zone	Back pain (3 mo)[c]	−0.1 (p = 0.4)[d]
High intensity zone[48]	No high intensity zone	Disability (3 mo)[c]	−0.2 (p = 0.1)[d]
Modic type 1[14]	No Modic	Back pain (3 mo)	−4.0 (−22.6 to 14.6)
Modic type 1[16]	No Modic	Back pain (3 mo)	5.0 (1.4 to 8.6)
Modic type 1[13]	No Modic	Back pain (3 mo)	−1.6 (−16.2 to 13.0)
Modic type 1[35]	No Modic	Back pain (6 mo)	7.0 (−5.3 to 19.3)
Modic type 1[15]	No Modic	Back pain (6 mo)	0.3 (−15.7 to 16.3)[e]
Modic type 1[14]	No Modic	Leg pain (3 mo)	−8.0 (−27.9 to 11.9)
Modic type 1[14]	No Modic	Disability (3 mo)	9.3 (−2.2 to 20.9)
Modic type 1[16]	No Modic	Disability (3 mo)	0.8 (−1.2 to 2.8)
Modic type 1[14]	Modic type 2	Back pain (3 mo)	−7.7 (−26.9 to 11.5)
Modic type 1[16]	Modic type 2	Back pain (3 mo)	−5.0 (−9.8 to −0.2)
Modic type 1[13]	Modic type 2	Back pain (3 mo)	−5.7 (−22.9 to 11.5)
Modic type 1[14]	Modic type 2	Leg pain (3 mo)	−18.4 (−43.9 to 7.1)
Modic type 1[14]	Modic type 2	Disability (3 mo)	−1.3 (−16.5 to 13.8)
Modic type 1[16]	Modic type 2	Disability (3 mo)	−5.0 (−7.5 to −2.5)
Modic type 1[35]	Modic type 2 and 3	Back pain (6 mo)	8.0 (−1.0 to 17.0)
Modic type 1 and 2[32]	No Modic	Back pain (1 mo)	7.3 (1.1 to 13.5)
Modic type 1 and 2[14]	No Modic	Back pain (3 mo)	0.7 (−13.2 to 14.6)
Modic type 1 and 2[16]	No Modic	Back pain (3 mo)	8.0 (4.5 to 11.5)
Modic type 1 and 2[13]	No Modic	Back pain (3 mo)	1.3 (−12.2 to 14.8)
Modic type 1 and 2[14]	No Modic	Leg pain (3 mo)	2.7 (−12.9 to 18.3)
Modic type 1 and 2[14]	No Modic	Disability (3 mo)	10.1 (0.3 to 20.0)
Modic type 1 and 2[16]	No Modic	Disability (3 mo)	3.8 (2.0 to 5.7)
Modic type 1 and/or 2[48]	No Modic	Back pain (3 mo)[c]	−0.3 (p = 0.2)d
Modic type 1 and/or 2[48]	No Modic	Disability (3 mo)[c]	0.1 (p = 0.5)d
Modic type 2[14]	No Modic	Back pain (3 mo)	3.7 (−11.7 to 19.1)
Modic type 2[16]	No Modic	Back pain (3 mo)	10.0 (6.0 to 14.0)
Modic type 2[13]	No Modic	Back pain (3 mo)	4.1 (−12.8 to 21.0)
Modic type 2[14]	No Modic	Leg pain (3 mo)	10.4 (−9.4 to 30.2)
Modic type 2[14]	No Modic	Disability (3 mo)	10.7 (−0.4 to 21.7)
Modic type 2[16]	No Modic	Disability (3 mo)	5.8 (3.8 to 7.8)
Modic type 2 and 3[35]	No Modic	Back pain (6 mo)	−1.0 (−8.1 to 6.1)
NRC[21]	No NRC	Back pain (4 mo)	−3.0 (−13.0 to 7.0)
NRC[29]	No NRC (all patients)	Disability (4 mo)	0.9 (0.4 to 2.0) (OR)
NRC[50]	No NRC	Disability (6 wk)	1.3 (0.7 to 2.5) (OR)
Spinal stenosis (severe)[50]	No spinal stenosis	Disability (6 wk)	0.4 (0.2 to 1.2) (OR)

Positive outcome for difference in means indicates exposure group had a worse outcome at follow-up compared with comparison group.
OR > 1 indicates greater odds of poor outcome in exposure group compared with comparison group.
Shaded rows indicate change in exposure or change in comparison MRI finding.
NR = not reported, NRC = nerve root compression.
[a] Short-term follow-up; outcome < 12 months (closest time point to 3 months) taken.
[b] Unless otherwise indicated.
[c] Average pain in the past 2 weeks.
[d] Beta coefficient.
[e] Change in pain scores.
[f] Z score.

changes and pain or disability outcomes at any follow-up time point. Two studies[22,28] found a relationship between greater increases in the volume of Modic type 1 changes between baseline and follow-up, and worse pain and disability outcomes over the same period (Table 4). Point estimates of the remaining four associations from two studies[24,44] suggested that there may be an association between those with Modic type 1 changes alone[24] or Modic type 1 and 2[44] changes and worse pain[24] or disability[44] outcomes, respectively, in the long term (Table 4).

Association between nerve root compression and pain and disability

Four studies[21,23,29,50] investigated the association between nerve root compression and pain[21,23] or disability[23,29,50] outcomes in populations with LBP. Pooling demonstrated no evidence of an association between the presence of nerve root compression compared with no compression and pain (MD −1.7, 95% CI −11.0 to 7.5)[21,23] or disability outcomes (OR 1.1, 95% CI 0.7 to 1.8)[29,50] in the long and short term, respectively. Results from associations based on two single studies[21,29] demonstrated no evidence of an association between the presence of nerve root compression compared with the absence of nerve root compression and pain outcomes in the short and long term[21] and disability outcomes in the long term[29] (Tables 3 and 4).

Association between disc degeneration and pain and disability

Eight studies[15,23–25,38,44,48,49] investigated 12 associations between disc degeneration and pain[15,23–25,48] or disability[23,38,44,48,49] outcomes in populations with LBP. Studies were only able to be pooled for two comparisons (which included four associations) [23,25,44,49] due to different measures of association reported. Pooling demonstrated that those with disc degeneration compared with no disc degeneration had slightly worse pain outcomes (MD 5.3, 95% CI 1.3 to 9.3)[23,25] and worse disability outcomes (OR 2.1, 95% CI 1.3 to 3.6)[44,49] in the long term. Point estimates of the remaining eight associations did not consistently suggest a relationship between those with disc degeneration compared with no disc degeneration and clinical outcomes; all of these associations had confidence intervals that included no association (or where they were not reported, were not statistically significant).

Table 4
Association between MRI findings and clinical outcomes in populations with current low back pain in the long term.

Exposure	Comparison	Clinical outcome (follow up)[a]	Difference in means (95% CI)[b]
Disc bulge[48]	No disc bulge	Back pain (1 y)[c]	0.19 (p = 0.15)[d]
Disc bulge[48]	No disc bulge	Disability (1 y)	0.22 (p = 0.11)[d]
Disc degeneration[48]	No disc degeneration	Back pain (1 y)[c]	−0.18 (p = 0.25)[d]
Disc degeneration (severe)[24]	No disc degeneration (severe)	Back pain (14 mo)	0.0 (p = 1.0)[d]
Disc degeneration[23]	No disc degeneration	Back pain (1 y)	8.1 (−1.2 to 18.1)
Disc degeneration[25]	No disc degeneration	Back pain (6 y)	4.7 (0.3 to 9.1)
Disc degeneration[23]	No disc degeneration	Disability (1 y)	10.4 (−9.6 to 29.7)
Disc degeneration[48]	No disc degeneration	Disability (1 y)[c]	−0.24 (p = 0.15)[d]
Disc degeneration[49]	No disc degeneration	Disability (18.5 mo)[e]	2.3 (1.3 to 4.0) (OR)
Disc degeneration[38]	No disc degeneration	Disability (13 y)	0.06 (p = 0.41)[d]
Disc degeneration[f,44]	No disc degeneration	Disability (2 y)	1.3 (0.3 to 5.3) (OR)
Disc extrusion[36]	Disc protrusion	Back pain (4 y)[g]	1.0 (0.6 to 1.6) (HR)
Disc extrusion[26]	Disc protrusion	Disability (8 y)	−2.4 (−5.2 to 0.4)[h]
Disc extrusion[36]	Disc protrusion	Leg pain (4 y)[f]	0.7 (0.5 to 1.2) (HR)
Disc extrusion[26]	Disc sequestration	Disability (8 y)	1.2 (−3.6 to 6.0)[h]
Disc extrusion (non-regression)[27]	Disc extrusion (partial regression)	Back pain (17 mo)	11.0 (−12.1 to 34.1)
Disc extrusion (non-regression)[27]	Disc extrusion (partial regression)	Disability (17 mo)	9.8 (−13.4 to 33.3)
Disc extrusion (non-regression)[27]	Disc extrusion (complete regression)	Back pain (17 mo)	18.0 (−2.1 to 38.1)
Disc extrusion (non-regression)[27]	Disc extrusion (complete regression)	Disability (17 mo)	14.2 (−4.9 to 33.3)
Disc extrusion (partial regression)[27]	Disc extrusion (complete regression)	Back pain (17 mo)	7.0 (−10.8 to 24.8)
Disc extrusion (partial regression)[27]	Disc extrusion (complete regression)	Disability (17 mo)	4.4 (−12.1 to 20.9)
Disc height reduction[23]	No disc height reduction	Back pain (1 y)	4.3 (−4.7 to 13.4)
Disc height reduction[25]	No disc height reduction	Back pain (6 y)	−4.8 (−10.2 to 0.5)
Disc height reduction[23]	No disc height reduction	Disability (1 y)	5.3 (−11.6 to 22.2)
Disc height reduction[f,44]	No disc height reduction	Disability (2 y)	1.4 (0.5 to 4.2) (OR)
Disc herniation[23]	No disc herniation	Back pain (1 y)	11.2 (−0.7 to 23)
Disc herniation[25]	No disc herniation	Back pain (6 y)	−7.0 (−16.3 to 2.4)
Disc herniation (large)[24]	No disc herniation (large)	Back pain (14 mo)	0.7 (p = 0.3)[d]
Disc herniation (posterolateral)[36]	Other disc herniation	Back pain (4 y)[g]	0.7 (0.4 to 1.1) (HR)
Disc herniation (posterolateral)[36]	Other disc herniation	Leg pain (4 y)[g]	0.6 (0.4 to 0.9) (HR)
Disc herniation (posterolateral)[26]	Disc herniation (other herniation location)	Disability (8 y)	−3.7 (−6.4 to −1.0)[h]
Disc herniation[23]	No disc herniation	Disability (1 y)	22.0 (−0.2 to 44.2)
Disc sequestration[36]	Disc protrusion	Back pain (4 y)[g]	0.6 (0.2 to 1.4) (HR)
Disc sequestration[26]	Disc protrusion	Disability (8 y)	−3.6 (−8.4 to 1.2)[h]
Disc sequestration[36]	Disc protrusion	Leg pain (4 y)[g]	0.9 (0.4 to 1.9) (HR)
Facet joint degeneration[38]	No facet joint degeneration	Disability (13 y)	−0.02 (p = 0.80)[d]
Facet joint arthropathy grade 2/3[f,44]	No facet joint arthropathy grade 2/3	Disability (2 y)	1.3 (0.3 to 5.3) (OR)
Lumbar spine osteoarthritis[23]	No lumbar spine osteoarthritis	Back pain (1 y)	−1.8 (−9.5 to 5.9)
Lumbar spine osteoarthritis[23]	No lumbar spine osteoarthritis	Disability (1 y)	−0.5 (−14.8 to 13.8)
High-intensity zone[48]	No high-intensity zone	Back pain (1 y)[c]	−0.38 (p = 0.01)[d]
High-intensity zone[23]	No high-intensity zone	Back pain (1 y)	4.2 (−4.2 to 12.6)
High-intensity zone[25]	No high-intensity zone	Back pain (6 y)	−0.3 (−4.3 to 3.8)
High-intensity zone[39]	No high-intensity zone	Disability (1 y)	−1.1 (?)
High-intensity zone[23]	No high-intensity zone	Disability (1 y)	5.0 (−9.5 to 19.4)
High-intensity zone[48]	No high-intensity zone	Disability (1 y)[c]	−0.08 (p = 0.59)[d]
High-intensity zone (posterior)[f,44]	No posterior high-intensity zone	Disability (2 y)	0.6 (0.2 to 1.8) (OR)
Hypertrophy ligamentous flava[25]	No hypertrophy ligamentous flava	Back pain (6 y)	8.4 (−3.7 to 20.4)
Modic type 1[14]	No Modic	Back pain (1 y)	−4.5 (−25.9 to 16.9)
Modic type 1[47]	No Modic	Back pain (1 y)[i]	2.0 (−6.2 to 10.3)
Modic type 1[35]	No Modic	Back pain (1 y)	1.0 (−12.0 to 14.0)
Modic type 1[23]	No Modic	Back pain (1 y)	−14.5 (−27.2 to −1.8)
Modic type 1[47]	No Modic	Back pain (1 y)[i]	1.0 (0.8 to 1.2) (HR)
Modic type 1[46]	No Modic	Back pain (14 mo)	1.0 (0.4 to 2.4) (OR)
Modic type 1[18]	No Modic	Back pain (1 y)	1.5 (0.1 to 15.9) (OR)
Modic type 1[k,24]	No Modic	Back pain (14 mo)	17.8 (2.6 to 33.0)[l]
Modic type 1 (large)[24]	No Modic	Back pain (14 mo)	0.1 (p = 0.9)[h]
Modic type 1 (large)[k,24]	No Modic	Back pain (14 mo)	−0.7 (−14.6 to 13.2)[l]
Modic type 1 (large)[j,24]	No Modic	Back pain (14 mo)	7.0 (−2.7 to 16.7)[l]
Modic type 1 (large)[m,24]	No Modic	Back pain (14 mo)	14.3 (0.8 to 27.9)[l]
Modic type 1[24]	No Modic	Back pain (14 mo)	−1.8 (p = 0.03)[d]
Modic type 1[j,24]	No Modic	Back pain (14 mo)	22.8 (12.2 to 33.4)[l]
Modic type 1[m,24]	No Modic	Back pain (14 mo)	27.8 (12.8 to 42.8)[l]
Modic type 1[25]	No Modic	Back pain (6 y)	−3.0 (−13.4 to 7.4)
Modic type 1[33]	No Modic	Back pain (10 y)	−5.0 (−23.6 to 13.6)
Modic type 1[14]	No Modic	Leg pain (1 y)	−4.8 (−24.8 to 15.2)
Modic type 1[33]	No Modic	Leg pain (10 y)	11.0 (−9.7 to 31.7)
Modic type 1[14]	No Modic	Disability (1 y)	2.9 (−10.6 to 16.5)
Modic type 1[47]	No Modic	Disability (1 y)	0.60 (−5.9 to 7.1)
Modic type 1[39]	No Modic	Disability (1 y)	−5.2 (−13.8 to 3.4)
Modic type 1[23]	No Modic	Disability (1 y)	−22.5 (−47.1 to 2.1)
Modic type 1[33]	No Modic	Disability (10 y)	−8.6 (−28.0 to 10.8)
Modic type 1[47]	No Modic	Disability (1 y)	1.00 (0.98 to 1.04) (HR)
Modic type 1[i,44]	No Modic	Disability (2 y)	0.7 (0.2 to 1.9) (OR)
Modic type 1[14]	Modic type 2	Back pain (1 y)	−1.3 (−28.4 to 25.8)
Modic type 1[18]	Modic type 2	Back pain (1 y)	1.0 (0.1 to 10.1) (OR)
Modic type 1[33]	Modic type 2	Back pain (10 y)	−8.0 (−26.6 to 10.6)
Modic type 1[14]	Modic type 2	Leg pain (1 y)	−18.9 (−43.5 to 5.7)
Modic type 1[33]	Modic type 2	Leg pain (10 y)	14.0 (−8.6 to 36.6)
Modic type 1[14]	Modic type 2	Disability (1 y)	−6.9 (−28.6 to 14.9)
Modic type 1[33]	Modic type 2	Disability (10 y)	4.0 (−20.2 to 28.2)

Table 4 (*Continued*)

Exposure	Comparison	Clinical outcome (follow up)[a]	Difference in means (95% CI)[b]
Modic type 1[35]	Modic type 2 and 3	Back pain (1 y)	0.0 (−10.5 to 10.5)
Modic type 1[33]	Modic type 3	Back pain (10 y)	18.0 (−19.1 to 55.1)
Modic type 1[33]	Modic type 3	Leg pain (10 y)	40.0 (−2.9 to 82.9)
Modic type 1[33]	Modic type 3	Disability (10 y)	38.0 (−6.6 to 82.6)
Modic type 1 change[28]	Not applicable as change measure	Back pain (1 y)	0.3 (NR)[o]
Modic type 1 change[22]	Not applicable as change measure	Back pain (2 y)	0.26 (p = 0.04)[d]
Modic type 1 change[28]	Not applicable as change measure	Disability (1 y)	0.3 (NR)[d]
Modic type 1 change[22]	Not applicable as change measure	Disability (2 y)	0.30 (p = 0.02)[a]
Modic type 1 and 2[14]	No Modic	Back pain (1 y)	−3.7 (−19.5 to 12.1)
Modic type 1 and 2[23]	No Modic	Back pain (1 y)	−18.0 (−29.2 to −6.8)
Modic type 1 and 2[18]	No Modic	Back pain (1 y)	1.4 (0.3 to 6.3) (OR)
Modic type 1 and 2[14]	No Modic	Leg pain (1 y)	7.1 (−8.0 to 22.2)
Modic type 1 and 2[14]	No Modic	Disability (1 y)	7.3 (−3.5 to 18.0)
Modic type 1 and 2[23]	No Modic	Disability (1 y)	−45.4 (−66.7 to 24.1)
Modic type 1 and 2[f,44]	No Modic type 1 and 2	Disability (2 y)	1.0 (0.3 to 3.5) (OR)
Modic type 1 or 2[f,44]	No Modic type 1 or 2	Disability (2 y)	3.4 (0.6 to 17.7) (OR)
Modic type 1 and/or 2[48]	No Modic	Back pain (1 y)[c]	0.17 (p = 0.26)[d]
Modic type 1 and/or 2[48]	No Modic	Disability (1 y)[c]	0.20 (p = 0.23)[d]
Modic type 1, 2, 3 (large)[24]	Modic type 1, 2, 3 (small)	Back pain (14 mo)	1.2 (p = 0.13)h
Modic type 1, 2, 3 (large)[k,24]	Modic type 1, 2, 3 (small)	Back pain (14 mo)	−12.3 (−27.4 to 2.8)[l]
Modic type 1, 2, 3 (large)[m,24]	Modic type 1, 2, 3 (small)	Back pain (14 mo)	2.5 (−12.2 to 17.2)[l]
Modic type 1, 2, 3 (large)[n,24]	Modic type 1, 2, 3 (small)	Back pain (14 mo)	−4.7 (−15.2 to 5.8)[l]
Modic type 1, 2, 3[33]	No Modic	Back pain (10 y)	−1.0 (−14.3 to 12.3)
Modic type 1, 2, 3[33]	No Modic	Leg pain (10 y)	0.0 (−15.4 to 15.4)
Modic type 1, 2, 3[33]	No Modic	Disability (10 y)	−13.2 (−28.3 to 1.9)
Modic type 1, 2, 3[38]	No Modic	Disability (13 y)	0.3 (NR) (R^2)
Modic type 1, 2, 3[38]	No Modic	Disability (13 y)	0.2 (p = 0.03)[d]
Modic type 1, 2 or 3[39]	No Modic	Disability (1 y)	3.8 (−3.0 to 10.7)
Modic type 2[14]	No Modic	Back pain (1 y)	−3.2 (−20.0 to 13.6)
Modic type 2[47]	No Modic	Back pain (1 y)[i]	1.0 (−5.1 to 7.6)
Modic type 2[23]	No Modic	Back pain (1 y)	−6.1 (−13.7 to 1.4)
Modic type 2[18]	No Modic	Back pain (1 y)	1.5 (0.3 to 7.1) (OR)
Modic type 2[47]	No Modic	Back pain (1 y)[i]	0.9 (0.8 to 1.1) (HR)
Modic type 2[33]	No Modic	Back pain (10 y)	3.0 (−12.2 to 18.2)
Modic type 2[14]	No Modic	Leg pain (1 y)	14.1 (−5.01 to 33.2)
Modic type 2[33]	No Modic	Leg pain (10 y)	−3.0 (−20.5 to 14.5)
Modic type 2[14]	No Modic	Disability (1 y)	9.8 (−0.5 to 20.0)
Modic type 2[47]	No Modic	Disability (1 y)	0.5 (−6.0 to 7.0)
Modic type 2[23]	No Modic	Disability (1 y)	−6.3 (−19.5 to 6.9)
Modic type 2[39]	No Modic	Disability (1 y)	0.2 (−5.1 to 5.5)
Modic type 2[33]	No Modic	Disability (10 y)	−12.6 (−29.7 to 4.5)
Modic type 2[f,44]	No Modic	Disability (2 y)	2.1 (0.7 to 5.9) (OR)
Modic type 2 and 3[35]	No Modic	Back pain (1 y)	1.0 (−7.1 to 9.1)
Modic type 2 or 3[25]	No Modic	Back pain (6 y)	1.2 (−4.1 to 6.4)
Modic type 2[33]	Modic type 3	Back pain (10 y)	26.0 (−11.4 to 63.4)
Modic type 2[33]	Modic type 3	Leg pain (10 y)	26.0 (−20.6 to 72.6)
Modic type 2[33]	Modic type 3	Disability (10 y)	34.0 (−16.4 to 84.4)
Modic type 2 change[28]	Not applicable as change measure	Back pain (1 y)	−0.2 (NR)[o]
Modic type 2 change[22]	Not applicable as change measure	Back pain (2 y)	−0.24 (p = 0.05)[d]
Modic type 2 change[28]	Not applicable as change measure	Disability (1 y)	−0.3 (NR)[o]
Modic type 2 change[22]	Not applicable as change measure	Disability (2 y)	−0.13 (p = 0.31)[d]
Modic type 3[33]	No Modic	Back pain (10 y)	−23.0 (−63.9 to 17.9)
Modic type 3[33]	No Modic	Leg pain (10 y)	−29.0 (−74.1 to 16.1)
Modic type 3[33]	No Modic	Disability (10 y)	−46.6 (−87.5 to −5.7)
NRC[21]	No NRC	Back pain (1 y)	−7.0 (−17.0 to 3.0)
NRC[23]	No NRC	Back pain (1 y)	2.5 (−5.5 to 10.4)
NRC[23]	No NRC	Disability (1 y)	2.9 (−11.1 to 16.8)
NRC[29]	No NRC (all patients)	Disability (1 y)	0.9 (0.5 to 2.0) (OR)
Schmorl lesion[25]	No Schmorl lesion	Back pain (6 y)	−3.6 (−7.6 to 0.4)
Spinal stenosis[23]	No spinal stenosis	Back pain (1 y)	2.7 (−6.9 to 12.2)
Spinal stenosis[25]	No spinal stenosis	Back pain (6 y)	−2.4 (−7.7 to 2.9)
Spinal stenosis[23]	No spinal stenosis	Disability (1 y)	−6.0 (−23.5 to 11.5)
Spondylolisthesis[25]	No spondylolisthesis	Back pain (6 y)	−16.6 (−42.4 to 9.1)
1 MRI abnormality[25]	No MRI abnormality	Back pain (6 y)	−0.6 (−5.4 to 4.2)
1 MRI abnormality[49]	No MRI abnormality	Back pain (18.5 mo)[e]	1.0 (0.4 to 3.1) (OR)
≥ 1 MRI abnormality[49]	No MRI abnormality	Back pain (18.5 mo)[e]	1.3 (0.5 to 3.0) (OR)
2 MRI abnormalities[25]	No MRI abnormality	Back pain (6 y)	−1.9 (−7.8 to 4.0)
2 MRI abnormalities[49]	No MRI abnormality	Back pain (18.5 mo)[e]	1.0 (0.4 to 2.7) (OR)
≥ 2 MRI abnormalities[49]	No MRI abnormality	Back pain (18.5 mo)[e]	1.3 (0.6 to 3.1) (OR)
3 MRI abnormalities[25]	No MRI abnormality	Back pain (6 y)	0.4 (−5.5 to 6.3)
3 MRI abnormalities[49]	No MRI abnormality	Back pain (18.5 mo)[e]	1.0 (0.4 to 2.7) (OR)
≥ 3 MRI abnormalities[49]	No MRI abnormality	Back pain (18.5 mo)[e]	1.5 (0.6 to 3.8) (OR)
4 MRI abnormalities[25]	No MRI abnormality	Back pain (6 y)	2.6 (−6.7 to 11.8)
4 MRI abnormalities[49]	No MRI abnormality	Back pain (18.5 mo)[e]	2.6 (1.0 to 7.1) (OR)
≥ 5 or more MRI abnormalities[25]	No MRI abnormality	Back pain (6 y)	−3.0 (−9.9 to 3.8)

Positive outcome for difference in means indicates exposure group had a worse outcome at follow-up compared with comparison group.
OR > 1 indicates greater odds of poor outcome in exposure group compared with comparison group.
HR > 1 indicates greater incidence of poor outcome in exposure group compared with comparison group.
Shaded rows indicate change in exposure or change in comparison MRI finding.
LBP = low back pain, MRI = magnetic resonance imaging, NRC = nerve root compression, RMDQ = Roland Morris Disability Questionnaire, R^2 = coefficient of determination, ? = the confidence interval reported in this study was implausible.

[a] Long-term follow-up; outcome ≥ 12 months (closest time point to 12 months) taken.
[b] Unless otherwise indicated.

[c] Average pain in the past 2 weeks.
[d] Beta coefficient.
[e] Disabling LBP in past 4 weeks (RMDQ > 11).
[f] Rehab group.
[g] Recurrence of pain.
[h] Change in disability score.
[i] Pain at rest during previous week.
[j] Rest + exercise group.
[k] Exercise group.
[l] Change in pain scores.
[m] Rest group.
[n] Rest + exercise group.
[o] Pearson's *r*.

Association between all other MRI findings and pain and disability

Thirteen studies[23–27,30,36,38,39,44,48–50] investigated 68 associations between the remaining 11 MRI findings (disc herniation,[23–27,30,36,50] disc bulge,[48] disc height reduction,[44,23,25] facet joint arthropathy,[44,38] high-intensity zone,[23,25,39,44,48] hypertrophy ligamentous flava,[25] lumbar osteoarthritis,[23] Schmorl lesion,[25] spinal stenosis,[23,25,50] spondylolisthesis,[25] one or multiple MRI abnormalities[25,49]) and clinical outcomes in populations with LBP. We were only able to pool results for five different comparisons (which included 11 associations from individual studies) in total (Table 5).

Pooling demonstrated no evidence of an association between disc height reduction (MD −1.0, 95% CI −9.8 to 7.9),[23,25] disc herniation (MD 6.0, 95% CI −9.3 to 21.3),[23,25,27] and spinal stenosis (MD −1.2, 95% CI −5.8 to 3.4)[23,25] compared with the absence of the respective MRI finding and pain outcomes in the long term. Pooling demonstrated no evidence of an association between the presence of a high-intensity zone compared with the absence of a high-intensity zone and pain (MD 0.6, 95% CI −3.1 to 4.2)[23,25] and disability outcomes (MD −0.4, 95% CI −5.3 to 4.6)[23,39] in the long term (Table 5).

We were unable to pool data for the remaining 57 associations. Point estimates in 44 of the 57 associations demonstrated no evidence of an association between those with an MRI finding at baseline and clinical outcomes at any follow-up time point, and 10 of the 57 associations suggested that those with an MRI finding at baseline may have worse clinical outcomes at any follow-up time point but had confidence intervals that included no association (or where confidence intervals were not reported, were not statistically significant) (Tables 3 and 4). The remaining three associations from single

Table 5
Pooled estimates for MRI findings and clinical outcomes for populations with current LBP and no LBP.

Exposure	Comparison	Clinical outcome (follow-up)	No. of participants	Difference in means (95% CI)[a]
Pooled estimates for Modic changes (regardless of type) compared with no Modic in populations with LBP				
Modic (type 1[13,14,15,16,35], type 2[13,14,16], and type 1 and 2[32])	No Modic	Back pain (ST)	852	6.6 (4.3 to 8.9)
Modic (type 1[14,16], type 2[14,16])	No Modic	Disability (ST)	273	4.6 (0.3 to 8.9)
Modic (type 1[14,23,24,25,47,35], type 2[14,23,47])	No Modic	Back pain (LT)	2,440	0.0 (−6.5 to 6.4)
Modic (type 1[14,23,47,39], type 2[14,23,47,39])	No Modic	Disability (LT)	784	−0.3 (−3.7 to 3.2)
Pooled estimates for MRI findings and clinical outcomes in populations with LBP				
Modic type 1[13,14,15,16,35]	No Modic	Back pain (ST)	308	4.4 (1.1 to 7.6)
Modic type 1[14,16]	No Modic	Disability (ST)	129	3.1 (−4.3 to 10.5)
Modic type 1[13,14]	No Modic	Improvement (ST)	95	1.5 (0.6 to 3.7) (OR)
Modic type 1[14,23,24,25,47,35]	No Modic	Back pain (LT)	2,171	1.2 (−9.1 to 11.4)
Modic type 1[14,23,47,39]	No Modic	Disability (LT)	334	−2.4 (−8.8 to 4.0)
Modic type 1[18,46]	No Modic	Back pain (LT)	140	1.1 (0.5 to 2.4) (OR)
Modic type 1 and 2[13,14,16,32]	No Modic	Back pain (ST)	624	7.2 (4.3 to 10.1)
Modic type 1 and 2[14,16]	No Modic	Disability (ST)	201	5.1 (0.2 to 10.0)
Modic type 1 and 2[13,23]	No Modic	Back pain (LT)	142	−12.0 (−25.8 to 1.8)
Modic type 1 and 2[13,23]	No Modic	Disability (LT)	142	−18.3 (−69.9 to 33.3)
Modic type 1 and 2[13,14,32]	No Modic	Improvement (ST)	495	1.5 (1.0 to 2.2) (OR)
Modic type 1[13,14,16]	Modic type 2	Back pain (ST)	165	−5.2 (−9.7 to −0.6)
Modic type 1[14,16]	Modic type 2	Disability (ST)	119	−4.9 (−7.4 to −2.5)
Modic type 1[13,14]	Modic type 2	Improvement (ST)	86	1.0 (0.4 to 2.6) (OR)
Modic type 2[13,14,16]	No Modic	Back pain (ST)	199	9.4 (5.6 to 13.3)
Modic type 2[14,16]	No Modic	Disability (ST)	144	5.9 (3.9 to 7.9)
Modic type 2[13,14]	No Modic	Improvement (ST)	101	1.5 (0.6 to 3.6) (OR)
Modic type 2[14,23,47]	No Modic	Back pain (LT)	269	−2.1 (−6.8 to 2.7)
Modic type 2[14,23,47,39]	No Modic	Disability (LT)	450	0.9 (−3.4 to 5.2)
NRC[21,23]	No NRC	Pain (LT)	437	−1.7 (−11.0 to 7.5)
NRC[29,50]	No NRC	Disability (ST)	598	1.1 (0.7 to 1.8) (OR)
Disc degeneration[23,25]	No disc degeneration	Pain (LT)	2,138	5.3 (1.3 to 9.3)
Disc degeneration[44,49]	No disc degeneration	Disability (LT)	305	2.1 (1.3 to 3.6) (OR)
Disc height reduction[23,25]	No disc height reduction	Pain (LT)	2,138	−1.0 (−9.8 to 7.9)
Disc herniation[23,25,27]	No disc herniation	Pain (LT)	2,172	6.0 (−9.3 to 21.3)
High-intensity zone[23,25]	No high-intensity zone	Pain (LT)	2,247	−0.6 (−3.1 to 4.2)
High-intensity zone[23,39]	No high-intensity zone	Disability (LT)	391	1.1 (−7.6 to 9.8)[b]
Spinal canal stenosis[23,25]	No spinal canal stenosis	Pain (LT)	2,138	−1.2 (−5.8 to 3.4)
Pooled estimates for MRI findings and clinical outcomes in populations with no LBP				
Disc degeneration[43,37]	No disc degeneration	Pain (LT)	90	1.9 (0.8 to 4.6) (OR)

Positive outcome for difference in means indicates exposure group had a worse outcome at follow-up compared with comparison group.
OR > 1 indicates greater odds of poor outcome in exposure group compared with comparison group.
Shaded rows indicate change in exposure or change in comparison MRI finding.
LT = long-term follow-up (outcome ≥ 12 months, closest time point to 12 months taken, NRC = nerve root compression, ST = short-term follow-up (outcome < 12 months, closest time point to 3 months taken).
[a] Unless otherwise indicated.
[b] Confidence intervals for Wilkens[39] for high-intensity zone were implausible therefore the standard deviation (SD) for Jensen[23] was used for the purpose of meta-analysis.

Table 6
Association between MRI findings and clinical outcomes in populations with no LBP.

Exposure	Comparison	Clinical outcome (follow-up duration)	HR (95% CI)[a]
Disc annular tear[20]	No disc annular tear	Back pain (1 y)[b]	1.0 (0.5 to 2.0)
Disc annular tear[19]	No disc annular tear	Back pain (1 y)[c]	1.2 (0.5 to 2.7)
Disc degeneration[25]	No disc degeneration	Back pain (6 ys)	0.3 (−0.1 to 0.7)[d]
Disc degeneration (worsening)[43]	Disc degeneration (no change)	Back pain (5 y)[e]	2.9 (0.8 to 10.3) (OR)
Disc degeneration Pfirrmann ≥ 3[20]	Disc degeneration Pfirrmann < 3	Back pain (1 y)[b]	2.6 (0.6 to 10.6)
Disc degeneration Pfirrmann ≥ 3[19]	Disc degeneration Pfirrmann < 3	Back pain (1 y)[c]	3.6 (0.5 to 26.2)
Disc degeneration[40]	No disc degeneration	Back pain (5 y)[e]	1.4 (0.4 to 4.6) (OR)
Disc degeneration[37]	No disc degeneration	Back pain (10 y)	1.3 (0.4 to 4.7) (OR)
Disc bulge[37]	No disc bulge	Back pain (10 y)	0.6 (0.2 to 2.4) (OR)
Disc height loss[25]	No disc height loss	Back pain (6 y)	0.2 (−0.4 to 0.7)[d]
Disc height loss[20]	No disc height loss	Back pain (1 y)[b]	1.3 (0.9 to 1.8)
Disc height loss[19]	No disc height loss	Back pain (1 y)[c]	3.3 (0.8 to 13.9)
Disc herniation[25]	No disc herniation	Back pain (6 y)	0.4 (−0.5 to 1.2)[d]
Disc herniation[40]	No disc herniation	Back pain (5 y)[e]	1.1 (0.3 to 4.2) (OR)
Disc herniation[20]	No disc herniation	Back pain (1 y)[b]	0.9 (0.4 to 2.3)
Disc herniation[19]	No disc herniation	Back pain (1 y)[c]	1.8 (0.4 to 7.5)
Disc protrusion[45]	No disc protrusion	Back and leg pain (3 y)[d]	0.5 (0.3 to 0.9)
Disc extrusion[45]	No disc extrusion	Back and leg pain (3 y)[d]	1.2 (0.4 to 3.4)
Disc high-intensity zone[25]	No disc high-intensity zone	Back pain (6 y)	0.3 (0.0 to 0.7)[d]
Disc high-intensity zone[20]	No disc high-intensity zone	Back pain (1 y)[b]	1.9 (1.0 to 3.5)
Disc high-intensity zone[19]	No disc high-intensity zone	Back pain (1 y)[c]	2.5 (1.2 to 5.1)
Facet joint arthrosis[20]	No facet joint arthrosis	Back pain (1 y)[b]	1.0 (0.5 to 1.9)
Facet joint osteoarthritis[19]	No facet joint osteoarthritis	Back pain (1 y)[c]	0.7 (0.3 to 1.5)
Hypertrophy ligamentous flava[25]	No hypertrophy ligamentous flava	Back pain (6 y)	−0.4 (−1.7 to 0.9)[d]
Modic type 1, 2, 3[20]	No Modic	Back pain (1 y)[b]	1.0 (0.5 to 2.3)
Modic type 1, 2, 3[19]	No Modic	Back pain (1 y)[c]	1.4 (0.6 to 3.4)
Modic type 1[25]	No Modic	Back pain (6 y)	0.1 (−0.8 to 1.1)[d]
Modic type 2 or 3[25]	No Modic	Back pain (6 y)	0.8 (−0.2 to 1.8)[d]
Nerve root contact[45]	No nerve root contact	Back and leg pain (3 y)[d]	1.9 (0.6 to 5.8)
Neural compromise[40]	No neural compromise	Back pain (5 y)[e]	2.3 (0.6 to 8.6) (OR)
Spinal canal stenosis[25]	No spinal canal stenosis	Back pain (6 y)	0.7 (0.0 to 1.4)[d]
Spinal canal stenosis[20]	No spinal canal stenosis	Back pain (1 y)[b]	0.7 (0.1 to 5.4)
Spinal central stenosis[45]	No spinal canal stenosis	Back and leg pain (3 y)[d]	1.9 (0.8 to 4.8)
Spondylolisthesis[25]	No spondylolisthesis	Back pain (6 y)	1.1 (−0.1 to 2.3)[d]
Spondylolisthesis[20]	No spondylolisthesis	Back pain (1 y)[b]	1.3 (0.6 to 2.8)
Spondylolisthesis[19]	No spondylolisthesis	Back pain (1 y)[c]	1.6 (0.7 to 3.6)
Schmorl lesion[25]	No Schmorl lesion	Back pain (6 y)	−0.5 (−0.9 to −0.2)[d]
Worsening abnormalities on MRI[5]	No worsening abnormalities on MRI	Back pain (7 y)	3.5 (NR) (RR)
1 MRI finding[25]	0 findings	Back pain (6 y)	0.3 (−0.1 to 0.6)[d]
1 finding present[19]	0 findings	Back pain (1 y)[c]	2.8 (0.4 to 22.0)
2 MRI findings[25]	0 findings	Back pain (6 y)	0.2 (−0.2 to 0.6)[d]
2 findings present[19]	0 findings	Back pain (1 y)[c]	5.9 (0.8 to 44.8)
3 findings present[19]	0 findings	Back pain (1 y)[c]	12.2 (1.3 to 118.2)
3 MRI findings[25]	0 findings	Back pain (6 y)	0.5 (−0.2 to 1.2)[d]
4 MRI findings[25]	0 findings	Back pain (6 y)	0.8 (0.0 to 1.5)[d]
≥ 5 MRI findings[25]	0 findings	Back pain (6 y)	1.2 (0.0 to 2.4)[d]

Positive outcome for difference in means indicates exposure group had a worse outcome at follow-up compared with comparison group.
HR > 1 indicates greater incidence of poor outcome in exposure group compared with comparison group.
OR > 1 indicates greater odds of poor outcome in exposure group compared with comparison group.
RR > 1 indicates that the poor outcome is more likely to develop in the exposure group.
Shaded rows indicate change in exposure or change in comparison MRI finding.
LBP = low back pain, MRI = magnetic resonance imaging, NR = not reported.
[a] Unless otherwise indicated.
[b] LBP (yes/no).
[c] Days to recurrence of activity limiting LBP.
[d] LBP recurrence.
[e] Beta coefficient.

studies demonstrated statistically significant results. The presence of a high-intensity zone compared with no high-intensity zone was associated with slightly better pain outcomes in the short term (β −0.38, p = 0.01).[48] The presence of disc herniation compared with no disc herniation was associated with worse disability outcomes (OR 2.7, 95% CI 1.4 to 5.1)[50] in the short term but slightly better disability outcomes (MD −3.7, 95% CI −6.4 to −1.0)[26] in the long term.

Association of MRI findings and clinical outcomes in people with no LBP

Eight studies[5,19,20,25,37,40,43,45] investigated 97 associations between MRI findings and pain outcomes in participants with no current LBP (Appendix 3). No studies reported on disability outcomes in populations with no current LBP. The MRI findings presented in three or more studies or where associations could be pooled across at least two studies are presented below. Results are presented in Table 6 and Appendix 3. Only one comparison (which included two associations) was able to be pooled. The remaining studies were unable to be pooled due to different outcome measures or measures of association reported across individual studies, or data from individual studies were based on the same study populations.[19,20,40,43]

Association between disc degeneration and pain

Six studies[19,20,25,37,40,43] reported on the association between disc degeneration and pain outcomes in populations with no LBP. Pooling demonstrated that the presence of disc degeneration compared with the absence of disc degeneration may increase the likelihood of experiencing pain in the long term (OR 1.9, 95% CI 0.8 to 4.6).[37,43] The remaining five studies[19,20,25,40,43] were unable to be pooled due to different outcome measures and measures of association having been used. Their associations suggested a consistent relationship of worse pain outcomes in those with disc degeneration compared with no or milder disc degeneration in the long term; only one study reported a confidence interval that excluded no association and even in this study the results were imprecise (OR 7.3, 95% CI 1.2 to 45.3)[43] (Appendix 3).

Table 7
Association between MRI findings and clinical outcome in mixed low back pain population samples.

Exposure	Comparison	Clinical outcome (follow-up duration)	OR (95% CI)[a]
Disc herniation[42]	No disc herniation	Back pain (5 y)[b]	4.81 (p = 0.01)
Disc degeneration (MRI group)[34]	No disc degeneration (MRI group)	Back pain (30 y)	−0.3 (−2.3 to 1.7) (MD)
Disc degeneration[41]	No disc degeneration	Back pain (5 y)[c]	4.40 (p = 0.08)
Disc degeneration[31]	No disc degeneration	Back pain (10 y)[d]	1.1 (1.1 to 1.1)
Disc degeneration (MRI group)[34]	No disc degeneration (MRI group)	Disability (30 y)	−2.9 (−14.0 to 8.2) (MD)
Disc degeneration[41]	No disc degeneration	Disability (5 y)[c]	1.1 (0.5 to 2.3)
Canal stenosis (moderate/severe)[41]	No canal stenosis	Back pain (5 y)[c]	2.90 (p = 0.09)
Canal stenosis (moderate/severe)[41]	No canal stenosis	Disability (5 y)[c]	2.1 (0.9 to 5.1)
Schmorl Node[31]	No Schmorl node	Back pain (10 y)[d]	1.1 (0.7 to 1.8)
Modic (moderate/severe)[41]	No Modic	Back pain (5 y)[c]	2.5 (p = 0.1)
Modic (type 1, 2 or 3)[31]	No Modic	Back pain (10 y)[d]	2.2 (1.4 to 3.3)

Positive outcome for difference in means indicates exposure group had a worse outcome at follow-up compared with comparison group.
OR > 1 indicates greater odds of poor outcome in exposure group compared with comparison group.
Shaded rows indicate change in exposure or change in comparison MRI finding.
LBP = low back pain, MRI = magnetic resonance imaging, NPRS = numerical pain rating scale.
[a] Unless otherwise indicated.
[b] Remissions of LBP for > 6 months.
[c] Serious LBP (NPRS ≥ 6 for at least 1 week).
[d] Disabling low back pain lasting > 1 month.

Association between disc herniation (all types) and pain

Five studies[19,20,25,40,45] reported on the association between the presence of disc herniation (including disc protrusion or extrusion) and pain outcomes in populations with no LBP. No studies were able to be pooled due to different outcome measures and measures of association used. Results of comparisons based on the five studies suggested no consistent relationship between the presence of disc herniation compared with no disc herniation and pain outcomes in the long term. No studies reported statistically significant associations (Table 6).

Association between all other MRI findings and pain

Five studies[5,19,20,25,37,40,45] investigated 36 associations between the remaining 13 MRI findings (disc annular tear,[19,20] disc bulge,[25,37] disc height loss,[19,20,25] high-intensity zone,[19,20,25] hypertrophy ligamentous flava,[25] facet joint osteoarthritis,[19,20] Modic type 1, 2 and/or 3 changes,[19,20,25] nerve root contact,[40,45] Schmorl lesion,[25] spinal canal stenosis,[20,25,45] spondylolisthesis,[19,20,25] worsening abnormalities on MRI,[5] one or more MRI finding[19,25]) and clinical outcomes in populations with no pain (Table 6). Point estimates from 26 of the 36 associations demonstrated no evidence of an association between those with an MRI finding at baseline and clinical outcomes, and eight of the 36 associations suggested that those with an MRI finding at baseline may have worse clinical outcomes at any follow-up time point (1 to 10 years). Only two associations from one study[19] reported associations between the presence of a high-intensity zone[19] or three aggregate MRI findings[19] compared with the absence of the respected MRI finding and worse pain outcomes at 1 year (OR 2.5, 95% CI 1.2 to 5.1 and OR 12.2, 95% CI 1.3 to 118.2, respectively)[19] (Table 6). However, results for three or more aggregate MRI findings[19] should be interpreted with caution due to the imprecise confidence intervals.

Association of MRI findings and clinical outcomes in mixed LBP population samples

Four studies[31,34,41,42] investigated 14 associations between MRI findings and clinical outcomes populations with mixed samples (Appendix 3); the results are presented in Table 7 and Appendix 3. No associations were able to be pooled due to different outcome measures or measures of association reported.

One study[31] demonstrated that the presence of Modic type 1, 2 or 3 changes compared with no Modic changes was associated with worse pain outcomes (OR 2.2, 95% CI 1.4 to 3.3) in the long term (10 years). One study[42] demonstrated that the presence of disc herniation compared with no disc herniation was associated with worse pain outcomes (OR 4.8, p = 0.01) in the long term (5 years) (Table 7 and Appendix 3). The remaining associations were not statistically significant. Point estimates in five of the 12 remaining associations suggested that those with an MRI finding at baseline may have worse clinical outcomes at any follow-up time point, and seven of the 12 remaining associations suggested no evidence of an association between those with an MRI finding at baseline and clinical outcomes at any follow-up time point.

Discussion

The available evidence is insufficient to judge whether most of the MRI findings that were tested were or were not associated with future pain and disability. The exceptions were Modic changes and disc degeneration, which were associated with slightly worse pain and disability outcomes. In populations with current LBP, the presence of Modic type 1 and 2 changes was associated with slightly worse pain and disability in the short term, the presence of Modic type 1 changes alone was associated with slightly worse pain in the short term, and the presence of disc degeneration was associated with slightly worse pain and worse disability outcomes in the long term. In populations with no LBP, pooled estimates demonstrated that the presence of disc degeneration may increase the likelihood of experiencing pain in the long term (Table 5).

This updated systematic review included 27 new studies and the 12 previous studies,[6] resulting in a total of 39 studies that investigated MRI findings and their relationship to future LBP. This review included a substantial number of studies (n = 39) in comparison with our previous review[6] (n = 12). Despite the increase in studies, pooling was limited due to the heterogeneity of the included studies. Studies varied in sample sizes (participants ranging from 20 to 1,997, median = 106) and used different outcome measures, different measures of associations, and different MRI measures. The results should be interpreted with caution, as they were based mostly on single studies, typically with small sample sizes and most associations did not demonstrate a statistically significant relationship between MRI findings and future LBP.

This review used the same sensitive search strategy that was previously used by Steffens et al[6] and followed a pre-specified protocol with clear inclusion criteria. A limitation of this review was the heterogeneity of the studies included, particularly the different outcome measures, MRI measures investigated and measures of associations reported. This limited the number of comparisons where a meta-analysis could be conducted, and the number of studies in the possible meta-analyses. However, the previous review was unable to perform any meta-analyses. Some studies included participants from RCTs[13–16,24,26,28,30,32,38,44,46,48,50] and it was not possible to determine the influence of the intervention on the relationships investigated; however, studies were excluded where the intervention was designed to change lumbar spine MRI findings (eg, surgery or antibiotic treatment for Modic changes).

A previous systematic review investigated the relationship between 11 different MRI findings and LBP in cross-sectional studies.[4]

Many MRI findings (disc herniation, disc degeneration, disc bulge, Modic type 1 changes and spondylolysis) were more prevalent in those with LBP (ORs ranging from 2.2, 95% CI 1.2 to 4.2 to 7.5, 95% CI 1.3 to 44.6).[4] Another review investigated cross-sectional associations between Modic changes and LBP.[59] A statistically significant association was found between Modic change (any type) and the presence of LBP in 15 of the 30 studies (ORs ranging from 1.5, 95% CI 1.0 to 2.3 to 83.1, 95% CI 4.9 to 1,424.1).[59] The findings from these two reviews cannot be directly compared with the current results as they investigated cross-sectional associations; however, the associations between Modic changes and pain in the cross-sectional studies[4,59] are consistent with the relationship to future LBP that the current review identified.

For those imaging findings (eg, Modic changes, disc degeneration) where some evidence of an association with future LBP was found, it is recommended that clinicians incorporate this information into their overall impression of a patient's likely prognosis when imaging findings are available, but imaging is not currently recommended purely to gain this prognostic information. For most MRI findings, it remains uncertain whether there is any relationship with future LBP. This review also suggests that the relationship between some imaging findings and future LBP may be different in people with and without LBP. For example, the presence of a high-intensity zone was associated with better outcomes in those with LBP but worse outcomes in those without LBP. This is important, as studies combining participants with and without LBP may miss a relevant association in one population. Our results should be interpreted with caution, as pooling mostly involved a small number of studies with varied sample sizes. There is still a paucity of large high-quality studies in this field of research.

Future research should focus on conducting high-quality studies to determine the clinical relevance of MRI findings and LBP. Analyses should investigate the impact of age and psychosocial factors on associations between MRI findings and clinical outcomes. Future studies should aim to standardise the assessment and reporting of clinical outcomes based on the recommended core outcome measures for trials in patients with non-specific LBP (eg, using the Oswestry Disability Index or Roland Morris Disability Questionnaire to measure physical functioning).[7] Standardising clinical outcome measures allows for better comparisons between study results and improves the interpretability of results in this area. Studies should also investigate the impact of the way clinical outcomes (eg, pain) and MRI findings are defined or categorised, as many studies continue to dichotomise clinical outcomes or MRI findings (present or absent) and do not account for different degrees of severity. This may result in missing important data or relationships. Future studies should also use standardised MRI protocols and sequences; MRI protocols should also include standardised descriptions and thresholds for positive MRI findings.

Our results suggest that some MRI findings may have weak associations with future LBP; however, for most of the MRI findings we evaluated it was unclear whether there was an association. Larger high-quality studies are still needed to resolve uncertainty for most MRI findings.

What was already known on this topic: The link between various findings on magnetic resonance imaging and low back pain is unclear because: structural changes on imaging can be present in both symptomatic and asymptomatic people; imaging findings become more prevalent with age; and much past research has used cross-sectional designs.

What this study adds: Based on longitudinal studies, some magnetic resonance imaging findings may have weak associations with future low back pain. The imaging findings that are predictive differ depending on whether pain is present at baseline. Although clinicians can incorporate this information into estimation of a patient's prognosis when imaging findings are available, it is not recommended to obtain imaging purely to gain this prognostic information.

Footnotes: [a] Comprehensive Meta-Analysis software Version 3, Biostat, Englewood, NJ, USA.

eAddenda: Appendices 1 to 3 can be found online at https://doi.org/10.1016/j.jphys.2023.02.007

Ethics approval: Not applicable.

Competing interests: The authors have no competing interests to declare.

Source(s) of support: Nil.

Acknowledgements: Nil.

Data sharing: The authors will make all relevant data available upon reasonable request.

Provenance: Not invited. Peer reviewed.

Correspondence: Christopher S Han, Institute for Musculoskeletal Health, The University of Sydney and Sydney Local Health District, Australia. Email: christopher.han@sydney.edu.au

References

1. Wu A, March L, Zheng X, Huang J, Wang X, Zhao J, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. *Ann Transl Med*. 2020;8:299.
2. van Tulder M, Becker A, Bekkering T, Breen A, del Real MT, Hutchinson A, et al. European guidelines for the management of acute nonspecific low back pain in primary care. *Eur Spine J*. 2006;15(Suppl 2):S169–S191.
3. Brinjikji W, Luetmer PH, Comstock B, Bresnahan BW, Chen LE, Deyo RA, et al. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *Am J Neuroradiol*. 2015;36:811–816.
4. Brinjikji W, Diehn FE, Jarvik JG, Carr CM, Kallmes DF, Murad MH, et al. MRI findings of disc degeneration are more prevalent in adults with low back pain than in asymptomatic controls: A systematic review and meta-analysis. *Am J Neuroradiol*. 2015;36:2394–2399.
5. Borenstein DG, O'Mara JW, Boden SD, Lauerman WC, Jacobson A, Platenberg C, et al. The value of magnetic resonance imaging of the lumbar spine to predict low-back pain in asymptomatic subjects. *J Bone Joint Surg Am*. 2001;83:1306–1311.
6. Steffens D, Hancock MJ, Maher CG, Williams C, Jensen TS, Latimer J. Does magnetic resonance imaging predict future low back pain? A systematic review. *Eur J Pain*. 2013;18:755–765.
7. Chiarotto A, Boers M, Deyo RA, Boers M, Buchbinder R, Corbin TP, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain*. 2018;159:481–495.
8. Altman DG. Systematic reviews of evaluations of prognostic variables. *BMJ*. 2001;323(7306):224–228.
9. da C Menezes Costa L, Maher CG, Hancock MJ, McAuley JH, Herbert RD, Costa LOP. The prognosis of acute and persistent low-back pain: a meta-analysis. *CMAJ*. 2012;184:E613–E624.
10. Pengel LHM, Herbert RD, Maher CG, Refshauge KM. Acute low back pain: systematic review of its prognosis. *BMJ*. 2003;327(7410):323.
11. Buttermann GR, Mullin WJ. Pain and disability correlated with disc degeneration via magnetic resonance imaging in scoliosis patients. *Eur Spine J*. 2008;17:240–249.
12. Mitra D, Cassar-Pullicino VN, McCall IW. Longitudinal study of vertebral type-1 end-plate changes on MR of the lumbar spine. *Eur J Radiol*. 2004;14:1574–1581.
13. Annen M, Peterson C, Humphreys BK. Comparison of treatment outcomes in nonspecific low-back pain patients with and without modic changes who receive chiropractic treatment. *J Manipulative Physiol Ther*. 2018;41:561–570.
14. Annen M, Peterson C, Leemann S, Schmid C, Anklin B, Humphreys BK. Comparison of outcomes in MRI confirmed lumbar disc herniation patients with and without modic changes treated with high velocity, low amplitude spinal manipulation. *J Manipulative Physiol Ther*. 2016;39:200–209.
15. Cai G, Laslett LL, Aitken D, et al. Effect of zoledronic acid and denosumab in patients with low back pain and modic change: A proof-of-principle trial. *J Bone Miner Res*. 2018;33:773–782.
16. Chen Y, Yang H, Zhang L, Wang Y, Zou J. Analyzing the influence of modic changes on patients with lower back pain undergoing conservative treatment. *Pain Res Manag*. 2019;2019:8185316.
17. de Schepper EIT, Koes BW, Oei EHG, Bierma-Zeinstra SMA, Luijsterburg PAJ. The added prognostic value of MRI findings for recovery in patients with low back pain in primary care: a 1-year follow-up cohort study. *Eur Spine J*. 2016;25:1234–1241.
18. el Barzouhi A, Vleggeert-Lankamp CLAM, van der Kallen BF, à Nijeholt GJ, van den Hout WB, Koes BW, et al. Back pain's association with vertebral end-plate signal changes in sciatica. *Spine*. 2014;14:225–233.
19. Hancock MJ, Kjaer P, Kent P, Jensen RK, Jensen TS. Is the number of different MRI findings more strongly associated with low back pain than single MRI findings? *Spine*. 2017;42:1283–1288.
20. Hancock MJ, Maher CM, Petocz P, Lin CW, Steffens D, Luque-Suarez A, et al. Risk factors for a recurrence of low back pain. *Spine*. 2015;15:2360–2368.
21. Harrisson SA, Ogollah R, Dunn KM, Foster NE, Konstantinou K. Prevalence, characteristics, and clinical course of neuropathic pain in primary care patients consulting with low back-related leg pain. *Clin J Pain*. 2020;36:813–824.
22. Järvinen J, Karppinen J, Niinimäki J, Haapea M, Grönblad M, Luoma K, et al. Association between changes in lumbar modic changes and low back symptoms over a two-year period. *BMC Musculoskelet Disord*. 2015;16:1–8.
23. Jensen OK, Nielsen CV, Sorensen JS, Stengaard-Pedersen K. Type 1 Modic changes was a significant risk factor for 1-year outcome in sick-listed low back pain

patients: a nested cohort study using magnetic resonance imaging of the lumbar spine. *Spine J*. 2014;14:2568–2581.
24. Jensen RK, Kent P, Hancock M. Do MRI findings identify patients with chronic low back pain and Modic changes who respond best to rest or exercise: a subgroup analysis of a randomised controlled trial. *Chiropr Man Therap*. 2015;23:1–9.
25. Kasch R, Truthmann J, Hancock MJ, Maher CG, Otto M, Nell C, et al. Association of lumbar MRI findings with current and future back pain in a population-based cohort study. *Spine*. 2022;47:201–211.
26. Kerr D, Zhao W, Lurie JD. What are long-term predictors of outcomes for lumbar disc herniation? A randomized and observational study. *Clin Orthop Relat Res*. 2015;473:1920–1930.
27. Kesikburun B, Eksioglu E, Turan A, Adiguzel E, Kesikburun S, Cakci A. Spontaneous regression of extruded lumbar disc herniation: Correlation with clinical outcome. *Pak J Med Sci*. 2019;35:974–980.
28. Koivisto K, Järvinen J, Karppinen J, Haapea M, Paananen M, Kyllönen E, et al. The effect of zoledronic acid on type and volume of modic changes among patients with low back pain. *BMC Musculoskelet Disord*. 2017;18:274.
29. Konstantinou K, Dunn KM, Ogollah R, Lewis M, van der Windt D, Hay EM, et al. Prognosis of sciatica and back-related leg pain in primary care: the ATLAS cohort. *Spine*. 2018;18:1030–1040.
30. Kuligowski T, Dębiec-Bak A, Skrzek A. Mobilisation efficacy in young patients with different stages of degenerative disc disease. *J Back Musculoskelet Rehabil*. 2020;33:913–918.
31. Määttä JH, Wadge S, MacGregor A, Karppinen J, Williams FMK. Vertebral endplate (modic) change is an independent risk factor for episodes of severe and disabling low back pain. *Spine*. 2015;40:1187–1193.
32. Peterson CK, Pfirrmann CWA, Hodler J. Are modic changes related to outcomes in lumbar disc herniation patients treated with imaging-guided lumbar nerve root blocks? *Eur J Radiol*. 2014;83:1786–1792.
33. Romero-Muñoz LM, Barriga-Martín A, Segura-Fragoso A, Martín-González C. Are Modic changes in patients with chronic low back pain indicative of a worse clinical course? 10 years of follow-up. *Rev Esp Cir Ortop Traumatol*. 2018;62:274–281.
34. Sääksjärvi S, Kerttula L, Luoma K, Paajanen H, Waris E. Disc degeneration of young low back pain patients: A prospective 30-year follow-up MRI study. *Spine*. 2020;45:1341–1347.
35. Schistad EI, Espeland A, Rygh LJ, Roe C, Gjerstad J. The association between Modic changes and pain during 1-year follow-up in patients with lumbar radicular pain. *Skeletal Radiol*. 2014;43:1271–1279.
36. Suri P, Pearson AM, Scherer EA, Zhao W, Lurie JD, Morgan TS, et al. Recurrence of pain after usual nonoperative care for symptomatic lumbar disk herniation: analysis of data from the spine patient outcomes research trial. *Phys Med Rehabil*. 2016;8:405–414.
37. Tonosu J, Oka H, Higashikawa A, Okazaki H, Tanaka S, Matsudaira K. The associations between magnetic resonance imaging findings and low back pain: A 10-year longitudinal analysis. *PLoS One*. 2017;12:e0188057.
38. Udby PM, Ohrt-Nissen S, Bendix T, Brorson S, Carreon LY, Andersen MO. The association of MRI findings and long-term disability in patients with chronic low back pain. *Global Spine J*. 2021;11:633–639.
39. Wilkens P, Scheel IB, Grundnes O, Hellum C, Storheim K. Prognostic factors of prolonged disability in patients with chronic low back pain and lumbar degeneration in primary care: a cohort study. *Spine*. 2013;38:65–74.
40. Boos N, Semmer N, Elfering A, Schade V, Gal I, Zanetti M, et al. Natural history of individuals with asymptomatic disc abnormalities in magnetic resonance imaging. *Spine*. 2000;25:1484–1492.
41. Carragee E, Alamin T, Cheng I, Franklin T, Hurwitz E. Does minor trauma cause serious low back illness? *Spine*. 2006;31:2942–2949.
42. Carragee EJ, Alamin TF, Miller JL, Carragee JM. Discographic, MRI and psychosocial determinants of low back pain disability and remission: a prospective study in subjects with benign persistent back pain. *Spine*. 2005;5:24–35.
43. Elfering A, Semmer N, Birkhofer D, Zanetti M, Hodler J, Boos N. Risk factors for lumbar disc degeneration: A 5-year prospective MRI study in asymptomatic individuals. *Spine*. 2002;27:125–134.
44. Hellum C, Johnsen LG, Gjertsen Ø, Berg L, Neckelmann G, Grundnes O, et al. Predictors of outcome after surgery with disc prosthesis and rehabilitation in patients with chronic low back pain and degenerative disc: 2-year follow-up. *Eur Spine J*. 2012;21:681–690.
45. Jarvik JG, Hollingworth W, Heagerty PJ, Haynor DR, Boyko EJ, Deyo RA. Three-year incidence of low back pain in an initially asymptomatic cohort: Clinical and imaging risk factors. *Spine*. 2005;30:1541–1548.
46. Jensen RK, Leboeuf-Yde C, Wedderkopp N, Sorensen JS, Jensen TS, Manniche C. Is the development of Modic changes associated with clinical symptoms? A 14-month cohort study with MRI. *Eur Spine J*. 2012;21:2271–2279.
47. Keller A, Boyle E, Skog TA, Cassidy JD, Bautz-Holter E. Are Modic changes prognostic for recovery in a cohort of patients with non-specific low back pain? *Eur Spine J*. 2012;21:418–424.
48. Kleinstück F, Dvorak J, Mannion AF. Are "structural abnormalities" on magnetic resonance imaging a contraindication to the successful conservative treatment of chronic nonspecific low back pain? *Spine*. 2006;31:2250–2257.
49. McNee P, Shambrook J, Harris EC, Kim M, Sampson M, Palmer KT, et al. Predictors of long-term pain and disability in patients with low back pain investigated by magnetic resonance imaging: a longitudinal study. *BMC Musculoskelet Disord*. 2011;12:234.
50. Modic MT, Obuchowski NA, Ross JS, Brant-Zawadzki MN, Grooff PN, Mazanec DJ, et al. Acute low back pain and radiculopathy: MR imaging findings and their prognostic role and effect on outcome. *Radiology*. 2005;237:597–604.
51. Beall DP, Wilson GL, Bishop R, Tally W. VAST Clinical Trial: Safely supplementing tissue lost to degenerative disc disease. *Int J Spine Surg*. 2020;14:239–253.
52. Berry DB, Rodriguez-Soto AE, Englund EK, Shahidi B, Parra C, Frank LR, et al. Multiparametric MRI characterization of level dependent differences in lumbar muscle size, quality, and microstructure. *Spine*. 2020;3:e1079.
53. Gornet MF, Bailey AS, Eltahawy HA, Sasso RC, Cheng WK, Palmer S, et al. 250. Interspinous process distraction compared to nonoperative care for moderate lumbar degenerative disc disease: results of the FDA IDE clinical trial. *Spine*. 2019;19:S122–S123.
54. Harrisson SA, Ogollah R, Dunn KM, Foster NE, Konstantinou K. Prognosis of low back-related leg pain patients with neuropathic pain: clinical course and prognostic factors. *Physiotherapy*. 2019;105:e10.
55. Mushlin AI. In patients with sciatica, MRI at 1 year did not differentiate between a favorable or unfavorable outcome. *Ann Transl Med*. 2013;159:JC12.
56. Brøgger HA, Maribo T, Christensen R, Schiøttz-Christensen B. Comparative effectiveness and prognostic factors for outcome of surgical and non-surgical management of lumbar spinal stenosis in an elderly population: protocol for an observational study. *BMJ open*. 2018;8:e024949–e.
57. Beall DP, Wilson GL, Gershon S, Davis TT, Bishop R, Tally WC, et al. 58. Viable disc tissue allograft supplementation in the treatment of degenerated intervertebral discs the one-year results of a randomized control trial. *Spine*. 2021;21:S29.
58. el Barzouhi A, Vleggeert-Lankamp CL, Lycklama à, Nijeholt GJ, Van der Kallen BF, van den Hout WB, Koes BW, et al. Influence of low back pain and prognostic value of MRI in sciatica patients in relation to back pain. *PloS One*. 2014;9:e90800.
59. Herlin C, Kjaer P, Espeland A, Skouen JS, Leboeuf-Yde C, Karppinen J, et al. Modic changes—their associations with low back pain and activity limitation: A systematic literature review and meta-analysis. *PloS One*. 2018;13:e0200677.

Appendix 1. Detailed search strategies.

EMBASE	
1	exp low back pain/ or low$ back pain$.mp.
2	exp backache/ or exp low back pain/
3	exp lumbago/
4	exp magnetic resonance imaging/
5	exp MRI/
6	exp magnetic resonance/
7	exp NMR/
8	exp nuclear magnetic resonance/
9	disc degen$.mp.
10	exp desiccation/
11	bulg$.mp.
12	exp protrusion/
13	extru$.mp.
14	nerve root$ compromis$.mp.
15	annular tear$.mp.
16	endplate chang$.mp.
17	exp stenosis/
18	facet degen$.mp.
19	high intens$ zone.mp.
20	modic chang$.mp.
21	degen$ disc disease.mp.
22	exp spondylolisthesis/
23	cohort analysis/
24	exp incidence/
25	prediction/ or predict$.mp.
26	(disease course or course).mp.
27	inception.mp.
28	survival/
29	logistic regression analysis/
30	exp proportional hazards model/ or cox.mp.
31	exp life tables/ or life table.mp.
32	exp log rank test/ or log rank.mp.
33	intervertebral disk degeneration/ or loss of disc height.mp. or intervertebral disk/
34	(prognosis or prognos$).mp.
35	Back Injuries/ or back injur*.mp.
36	follow up/
37	1 or 2 or 3 or 35
38	4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 33
39	23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 34 or 36
40	37 and 38 and 39
41	limit 40 to human
42	limit 41 to (child or preschool child <1 to 6 years> or school child <7 to 12 years> or adolescent <13 to 17 years> or adult <18 to 64 years> or aged <65+ years>)
43	limit 42 to yr="2012 -Current"

MEDLINE	
1	exp Cohort Studies/
2	Incidence/
3	follow-up stud$.mp.
4	prognos$.mp.
5	predict$.mp.
6	cours$.mp.
7	incept$.mp.
8	surviv$.mp.
9	logistic$.mp.
10	cox.mp.
11	life tabl$.mp.
12	log rank.mp.
13	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12
14	low$ back pain$.mp.
15	back pain$.mp.
16	lumbago.mp.
17	back injur$.mp.
18	backache.mp.
19	14 or 15 or 16 or 17 or 18
20	magnetic resonance imaging/
21	MRI$.mp.
22	(magnetic adj5 resonance).mp.
23	NMR.mp.
24	nuclear magnetic resonance.mp.
25	disc degen$.mp.
26	desiccat$.mp.
27	loss$ of disc$ height$.mp.
28	bulg$.mp.
29	protru$.mp.
30	extru$.mp.
31	nerve root compromise.mp.
32	annular tear$.mp.
33	endplate chang$.mp.
34	steno$.mp.
35	facet degen$.mp.
36	high intensity zone$.mp.
37	modic chang$.mp.
38	degen$ disc$ diseas$.mp.
39	spondylo$.mp.
40	20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39
41	13 and 19 and 40
42	limit 41 to yr="2012 -Current"
43	limit 42 to (humans and ("all child (0 to 18 years)" or "all adult (19 plus years)"))

CINAHL	
1	MH log-rank test
2	"life table method*"
3	MH logistic regression+
4	MH survival analysis+ OR MH cox proportional hazards model
5	"incept*"
6	"cours*"
7	MH prognosis+
8	"predic*"
9	MH incidence+
10	MH prospective studies+
11	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10
12	"backach*"
13	(MH "Back Injuries+")
14	"lumbago"
15	MH low back pain OR MH back pain+
16	12 or 13 or 14 or 15
17	11 AND 16
18	MH Spondylolisthesis+
19	"Degen* disc disease"
20	"modic chang*"
21	"high intens* zone"
22	"facet degen*"
23	"stenos*"
24	"endplate chang*"
25	"annular tear*"
26	"nerve root* compromis*"
27	"extru*"
28	"protru*"
29	"bulg*"
30	"los* of disc* height"
31	"Desicc*"
32	"disc degen*"
33	MH nri OR MH Magnetic Resonance Spectroscopy
34	"MRI OR magnetic resonance"
35	18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34
36	17 AND 35
37	Limiters - Published Date: 20120101-20211231; Age Groups: Child: 6-12 years, Adolescent: 13-18 years, Adult: 19-44 years, Middle Aged: 45-64 years, Aged: 65+ years, Aged, 80 and over

Appendix 2. List of excluded full-text articles and the primary reason for exclusion.

#	Study	Title	First reason for exclusion
1	Acosta et al 2021	Radiographic evaluation of low back pain after lumbar fusion using Technetium-99 single-photon emission computed tomography: Impact on clinical decision-making and outcome	Not a prospective study
2	Albert et al 2013	Antibiotic treatment in patients with chronic low back pain and vertebral bone oedema (Modic type 1 changes): a double-blind randomized clinical controlled trial of efficacy	No association
3	Al-Kaisy et al 2020	10 kHz spinal cord stimulation for the treatment of non-surgical refractory back pain: subanalysis of pooled data from two prospective studies	No comparison MRI finding used
4	Altun et al 2017	Lumbar herniated disc: spontaneous regression	No LBP outcome at follow-up
5	Ansari et al 2020	Outcome of physical rehabilitation in patients indicated for surgery for chronic low back pain	Abstract only
6	Banno et al 2021	Clinical outcome of condoliase injection treatment for lumbar disc herniation: Indications for condoliase therapy	Not a prospective study
7	Basques (a) et al 2017	The Kinematics and Spondylosis of the Lumbar Spine Vary Depending on the Levels of Motion Segments in Individuals with Low Back Pain	No LBP outcome at follow-up
8	Basques (b) et al 2016	The Kinematics and Spondylosis of the Lumbar Spine Vary Depending on the Levels of Motion Segments in Individuals with Low Back Pain	No LBP outcome at follow-up
9	Beall et al 2019	Intervertebral disc regeneration techniques	Abstract only
10	Beall et al 2020	VAST clinical trial: Safely supplementing tissue lost to degenerative disc disease	Abstract only
11	Berg et al 2020	The prevalence of lumbar disc degeneration in symptomatic younger patients: A study of MRI scans	No LBP outcome at follow-up
12	Berry et al 2019	Multiparametric MRI characterization of level dependent differences in lumbar muscle size, quality, and microstructure	No LBP outcome at follow-up
13	Bianchi et al 2015	Are the presence of MODIC changes on MRI scans related to “improvement” in low back pain patients treated with lumbar facet joint injections?	No MRI at baseline
14	Biassoni et al 2014	Bone scintigraphy with SPECT CT in paediatric patients with back pain: Initial experience of a tertiary paediatric referral centre	Abstract only
15	Biassoni et al 2018	Bone scintigraphy with SPECT CT in paediatric patients with back pain: The experience of a paediatric tertiary referral centre	Abstract only
16	Boody et al 2020	Evaluation of DIAMTM Spinal Stabilization System for lower lumbar disc degenerative disease: A randomized, prospective, single-site study	No association
17	Braten et al 2019	Efficacy of antibiotic treatment in patients with chronic low back pain and Modic changes (the AIM study):	No association

		double blind, randomised, placebo controlled, multicentre trial	
18	Brauer et al 2020	Occupational lifting predicts hospital admission due to low back pain in a cohort of airport baggage handlers	No MRI at baseline
19	Brogger et al 2018	Comparative effectiveness and prognostic factors for outcome of surgical and non-surgical management of lumbar spinal stenosis in an elderly population: protocol for an observational study	Protocol only
20	Brooks et al 2018	Back Pain in a Pediatric Emergency Department: Etiology and Evaluation	No LBP outcome at follow-up
21	Burgstaller et al 2016	Is There an Association Between Pain and Magnetic Resonance Imaging Parameters in Patients with Lumbar Spinal Stenosis?	Not a prospective study
22	Burgstaller et al 2020	Long-term Results After Surgical or Nonsurgical Treatment in Patients with Degenerative Lumbar Spinal Stenosis	No LBP outcome at follow-up
23	Buttermann et al 2008	Pain and disability correlated with disc degeneration via magnetic resonance imaging in scoliosis patients	No LBP outcome at follow-up
24	Celenlioglu et al 2019	Does facet tropism negatively affect the response to transforaminal epidural steroid injections? A prospective clinical study	Not a prospective study
25	Chandra et al 2016	An algorithmic approach for clinical management of low back pain	No MRI at baseline
26	Chao et al 2017	Association between menopause and lumbar disc degeneration: an MRI study of 1,566 women and 1,382 men	Not a prospective study
27	Chen et al 2014	Radiological signs of Scheuermann's disease and low back pain: Lumbar MRI classification and six-year low back pain follow-up in 188 hospital staff members	Abstract - Found full text which is Liu et al 2014
28	Chen et al 2021	Association of Lumbar Spine Radiographic Changes with Severity of Back Pain–Related Disability Among Middle-aged, Community-Dwelling Women	MRI not used
29	Chepurin et al 2019	Bony stress in the lumbar spine is associated with intervertebral disc degeneration and low back pain: a retrospective case–control MRI study of patients under 25 years of age	No LBP outcome at follow-up
30	Cox et al 2017	Lumbar magnetic resonance imaging findings in patients with and without Waddell Signs	No LBP outcome at follow-up
31	Cuellar et al 2016	Does provocative discography cause clinically important injury to the lumbar intervertebral disc? A 10-year matched cohort study	No association
32	Davis et al 2020	Abstract only	Abstract only
33	de Schepper et al 2016	Prevalence of spinal pathology in patients presenting for lumbar MRI as referred from general practice	Not a prospective study
34	Demirel et al 2017	Regression of lumbar disc herniation by physiotherapy. Does non-surgical spinal decompression therapy make a difference? Double-blind randomized controlled trial	No comparison MRI finding used
35	Dessouky et al 2018	Magnetic Resonance Neurography in Chronic Lumbosacral and Pelvic Pain: Diagnostic and Management Impact: An Institutional Audit	No MRI at baseline

36	Dragsbaek 2020	An exploratory study of different definitions and thresholds for lumbar disc degeneration assessed by MRI and their associations with low back pain using data from a cohort study of a general population	Not prospective
37	Eksi et al 2020	Lumbar intervertebral disc degeneration, end-plates and paraspinal muscle changes in children and adolescents with low-back pain	Not a prospective study
38	el Barzouhi et al 2014	Influence of low back pain and prognostic value of MRI in sciatica patients in relation to back pain	Not enough data for association
39	Elkholy et al 2019	Spontaneous Resorption of Herniated Lumbar Disk: Observational Retrospective Study in 9 Patients	No association
40	Fritz et al 2016	Chronic back pain is associated with decreased prefrontal and anterior insular gray matter. Results from a population-based cohort study	No MRI at baseline
41	Furunes et al 2018	Adjacent Disc Degeneration After Lumbar Total Disc Replacement or Nonoperative Treatment	No LBP outcome at follow-up
42	Goksenoglu et al 2021	Predictors of functional mobility in patients with symptomatic degenerative lumbar spinal stenosis	Not a prospective study
43	Gornet et al 2019	250. Interspinous process distraction compared to nonoperative care for moderate lumbar degenerative disc disease: results of the FDA IDE clinical trial	Abstract only
44	Gu et al 2019	A 10-year follow-up of electric acupuncture for lumbar disc herniation and sciatica	No association
45	Gupta et al 2020	Does Size Matter? An Analysis of the Effect of Lumbar Disc Herniation Size on the Success of Nonoperative Treatment	No LBP outcome at follow-up
46	Hadzic et al 2013	Low back and lumbar radicular syndrome: Comparative study of the operative and non-operative treatment	MRI not used
47	Hanimoglu et al 2019	Effects of Modic Type 1 Changes in the Vertebrae on Low Back Pain	No LBP outcome at follow-up
48	Harrison et al 2019	Prognosis of low back-related leg pain patients with neuropathic pain: clinical course and prognostic factors	Abstract only
49	Hashizume et al 2020	Viable allograft for intervertebral disc supplementation-provisional results vast trial	Abstract only
50	Hong et al 2016	Resolution of lumbar disk herniation without surgery	Not a prospective study
51	Jain et al 2020	Effect of multimodal physical rehabilitation protocol on walking and standing time in patients with leg pain in lumbar degenerative spondylolisthesis	Abstract only
52	Jamaludin et al 2019	Disc degeneration: More than an aging process	Abstract only
53	Jang et al 2016	Lumbar intervertebral disc degeneration and related factors in Korean firefighters	Not a prospective study
54	Jensen et al 2012	Rest versus exercise as treatment for patients with low back pain and Modic changes. A randomized controlled clinical trial	No association
55	Jensen et al 2019	Probiotics for chronic low back pain with type 1 Modic changes: a randomized double-blind, placebo-controlled trial with 1-year follow-up using *Lactobacillus Rhamnosis GG*	No comparison MRI finding used

56	Jin et al 2021	Predictive Classification System for Low Back Pain Based on Unsupervised Clustering	No LBP outcome at follow-up
57	Jung et al 2020	A prospective study of non-surgical versus surgical treatment for lumbar spinal stenosis without instability	No comparison MRI finding used
58	Karademir et al 2017	Adolescent lumbar disc herniation: Impact, diagnosis, and treatment	No comparison MRI finding used
59	Kigozi et al 2019	Factors associated with costs and health outcomes in patients with Back and leg pain in primary care: a prospective cohort analysis	No LBP outcome at follow-up
60	Kim et al 2013	The Influence of Pain Sensitivity on the Symptom Severity in Patients with Lumbar Spinal Stenosis	No LBP outcome at follow-up
61	Kim et al 2015	Clinical symptoms of lumbar spinal stenosis associated with morphological parameters on magnetic resonance images with morphological parameters on magnetic resonance images	No MRI at baseline
62	Kohat et al 2017	Clinical significance of magnetic resonance imaging findings in chronic low backache	No association
63	Kojima et al 2018	Lumbar intervertebral disc degeneration in professional surfers	Not a prospective study
64	Kozaki et al 2020	Fatty infiltration in lumbar paravertebral muscle is a risk factor for low back pain and related disorders: A 3 years' follow-up of the general population in Japan	Abstract only
65	Kristoffersen et al 2021	Oedema on STIR modified the effect of amoxicillin as treatment for chronic low back pain with Modic changes—subgroup analysis of a randomized trial	No association
66	Kristoffersen et al 2021	Oedema on STIR modified the effect of amoxicillin as treatment for chronic low back pain with Modic changes—subgroup analysis of a randomized trial	No LBP outcome at follow-up
67	Lee et al 2017	Long-Term Course to Lumbar Disc Resorption Patients and Predictive Factors Associated with Disc Resorption	No association
68	Leeman et al 2014	Outcomes of acute and chronic patients with magnetic resonance imaging–confirmed symptomatic lumbar disc herniations receiving high-velocity, low-amplitude, spinal manipulative therapy: a prospective observational cohort study with one-year follow-up	No association
69	Lindsay et al 2021	Assessing need for advanced imaging among athletic children with chronic back pain using a numerical rating scale	Abstract only
70	Liu et al 2014	Radiological Signs of Scheuermann Disease and Low Back Pain	No LBP outcome at follow-up
71	Luoma et al 2016	Chronic low back pain in relation to Modic changes, bony endplate lesions, and disc degeneration in a prospective MRI study	No association
72	Maurer et al 2020	Long-term effect of physical inactivity on thoracic and lumbar disc degeneration—an MRI-based analysis of 385 individuals from the general population	No association
73	McKeag et al 2020	Assessment of the utility of the National Health Service England Low Back and Radicular Pain Pathway: analysis of patient reported outcomes	No comparison MRI finding used

74	Mera et al 2021	Association between types of Modic changes in the lumbar region and low back pain in a large cohort: the Wakayama spine study	Not prospective
75	Mera et al 2021	Association between types of Modic changes in the lumbar region and low back pain in a large cohort: the Wakayama spine study	Not a prospective study
76	Minetama et al 2018	A comparative study of 2-year follow-up outcomes in lumbar spinal stenosis patients treated with physical therapy alone and those with surgical intervention after less successful physical therapy	No comparison MRI finding used
77	Minetama et al 2020	Therapeutic Advantages of Frequent Physical Therapy Sessions for Patients with Lumbar Spinal Stenosis	No association
78	Mitra et al 2004	Longitudinal study of vertebral type-1 end-plate changes on MR of the lumbar spine	No comparison MRI finding used
79	Monro et al 2015	Disc extrusions and bulges in nonspecific low back pain and sciatica: Exploratory randomised controlled trial comparing yoga therapy and normal medical treatment	No association
80	Mukasa et al 2020	A prediction model of low back pain risk: a population-based cohort study in Korea	No MRI at baseline
81	Munir et al 2018	Endplate Defect Is Heritable, Associated with Low Back Pain and Triggers Intervertebral Disc Degeneration	No association
82	Nolte et al 2020	P42. When does MRI change management of pediatric back pain? A predictive scoring system	Abstract only
83	Obanife et al 2021	Management outcomes of lumbar spine degenerative diseases: Comparing operative versus non-operative treatments using Swiss spine stenosis scoring	No MRI at baseline
84	Ohashi et al 2018	Predicting Factors at Skeletal Maturity for Curve Progression and Low Back Pain in Adult Patients Treated Nonoperatively for Adolescent Idiopathic Scoliosis with Thoracolumbar/Lumbar Curves	No comparison MRI finding used
85	Oktay et al 2019	Spontaneous Regression of Lumbar Disc Herniations: A Retrospective Analysis of 5 Patients	Not a prospective study
86	Panta et al 2015	Morphological Changes in Degenerative Disc Disease on Magnetic Resonance Imaging: Comparison Between Young and Elderly	No LBP outcome at follow-up
87	Park et al 2019	Predictors of Response to a Medial Branch Block: MRI Analysis of the Lumbar Spine	No LBP outcome at follow-up
88	Peng et al 2012	Prospective Clinical Study on Natural History of Discogenic Low Back Pain at 4 Years of Follow-up	No MRI at baseline
89	Peterson et al 2013	Symptomatic magnetic resonance imaging–confirmed lumbar disk herniation patients: a comparative effectiveness prospective observational study of 2 age- and sex-matched cohorts treated with either high-velocity, low-amplitude spinal manipulative therapy or imaging-guided lumbar nerve root injections	No comparison MRI finding used
90	Peul et al 2013	In patients with sciatica, MRI at 1 year did not differentiate between a favorable or unfavorable outcome	Abstract only
91	Raudner et al 2019	Prediction of Lumbar Disk Herniation and Clinical Outcome Using Quantitative Magnetic Resonance Imaging A 5-Year Follow-Up Study	No outcome measure used

92	Sairyo et al 2016	State-of-the-art management of low back pain in athletes: Instructional lecture	Not a prospective study
93	Schistad et al 2019	Five-year development of lumbar disc degeneration—a prospective study	No LBP outcome at follow-up
94	Schneider et al 2016	Exploratory analysis of clinical predictors of outcomes of nonsurgical treatment in patients with lumbar spinal stenosis	No LBP outcome at follow-up
95	Schroeder et al 2016	The role of intense athletic activity on structural lumbar abnormalities in adolescent patients with symptomatic low back pain	No LBP outcome at follow-up
96	Shimozaki et al 2018	Incidence rates and characteristics of abnormal lumbar findings and low back pain in child and adolescent weightlifter: A prospective three-year cohort study	No association
97	Shin et al 2014	The long-term course of patients undergoing alternative and integrative therapy for lumbar disc herniation: 3-year results of a prospective observational study	Not enough data for association
98	Shin et al 2016	Long-Term Course of Alternative and Integrative Therapy for Lumbar Disc Herniation and Risk Factors for Surgery	Not enough data for association
99	Shokri et al 2018	Spinal manipulation in the treatment of patients with MRI-confirmed lumbar disc herniation and sacroiliac joint hypomobility: a quasi-experimental study	Not a prospective study
100	Smith et al 2021	The Association Between Different Trajectories of Low Back Pain and Degenerative Imaging Findings in Young Adult Participants within The Raine Study	Not prospective
101	Smith et al 2014	Operative and Nonoperative Treatment Approaches for Lumbar Degenerative Disc Disease Have Similar Long-Term Clinical Outcomes Among Patients with Positive Discography	MRI not used
102	Smuck et al 2021	Prospective, randomized, multicenter study of intraosseous basivertebral nerve ablation for the treatment of chronic low back pain: 12-month results	No comparison MRI finding used
103	Son et al 2021	Disc height discrepancy between supine and standing positions as a screening metric for discogenic back pain in patients with disc degeneration	No LBP outcome at follow-up
104	Suri et al 2012	Recurrence of Radicular Pain or Back Pain After Nonsurgical Treatment of Symptomatic Lumbar Disk Herniation	No association reported (Leg pain not LBP)
105	Suri et al 2014	Longitudinal associations between incident lumbar spine MRI findings and chronic low back pain or radicular symptoms: retrospective analysis of data from the longitudinal assessment of imaging and disability of the back (LAIDBACK)	No appropriate outcome measure
106	Svanbergsson et al 2017	Magnetic resonance imaging in the diagnosis of lumbar spine pain: Utilization, relationship to symptoms and effects on treatment	No MRI at baseline
107	Tavares et al 2021	Intradiscal glucocorticoids injection in chronic low back pain with active discopathy: A randomized controlled study	No MRI at baseline

108	Toyooka et al 2019	High incidence rate of lumbar spinal disease among child and adolescent weightlifting athletes: A prospective 4-year cohort study	Abstract only
109	Udby et al 2019	Modic Changes Are Not Associated with Long-term Pain and Disability: A Cohort Study With 13-year Follow-up	Group received surgery
110	Udby et al 2019	148. Which MRI findings are associated with long-term disability in low back pain patients?	Abstract only
111	Udby et al 2020	The Association of MRI Findings and Long- Term Disability in Patients with Chronic Low Back Pain	No association
112	Vagaka et al 2019	Do lumbar magnetic resonance imaging changes predict neuropathic pain in patients with chronic non-specific low back pain?	Not a prospective study
113	van den Berg et al 2019	Lumbar disc degeneration is longitudinally associated with low back pain presence in older adults	Abstract only
114	van den Berg et al 2020	Clinical and radiographic features of spinal osteoarthritis predict long-term persistence and severity of back pain in older adults	MRI not used
115	Wang et al 2018	Factors associated with lumbar disc high- intensity zone (HIZ) on T2-weighted magnetic resonance image: a retrospective study of 3185 discs in 637 patients	No LBP outcome at follow-up
116	Yamada et al 2014	Targeted therapy of low back pain associated with de novo degenerative lumbar scoliosis in the elderly: Prospective observation cohort study	Abstract only
117	Yamada et al 2021	Long-term outcome of targeted therapy for low back pain in elderly degenerative lumbar scoliosis	No comparison MRI finding used
118	Yang et al 2016	The association between back pain and level of lumbar spine according to Modic changes among Korean	Abstract only
119	Ye et al 2015	Comparison of lumbar spine stabilization exercise versus general exercise in young male patients with lumbar disc herniation after 1 year of follow-up	Not a prospective study
120	Yokoe et al 2021	Predictors of Spondylolysis on Magnetic Resonance Imaging in Adolescent Athletes with Low Back Pain	No MRI at baseline
121	Yu et al 2016	The Use of Lumbar Spine Magnetic Resonance Imaging in Eastern China: Appropriateness and Related Factors	No LBP outcome at follow-up
122	Zhao et al 2020	Long-Term Outcomes of Epidurals with Lidocaine with or Without Steroids for Lumbar Disc Herniation and Spinal Stenosis: A Meta-Analysis	Not a prospective study
123	Zhou et al 2018	Clinical efficacy of calcitonin compared to diclofenac sodium in chronic nonspecific low back pain with type I Modic changes: a retrospective study	No comparison MRI finding used

LBP = low back pain, MRI = magnetic resonance imaging

Appendix 3. All associations between MRI findings and clinical outcomes in all populations.

Study	Exposure	Comparison	Clinical outcome (follow-up duration)	Difference in means (95% CI) unless indicated
			Populations with current LBP	
Kleinstuck et al 2006	Disc bulge	No disc bulge	Average pain in the past 2 wks (VAS, 0 to 10) (3 mo)	0.08 (p = 0.60)[a,b]
Kleinstuck et al 2006	Disc bulge	No disc bulge	Average pain in the past 2 wks (VAS, 0 to 10) (12 mo)	0.19 (p = 0.15) [a,b]
Kleinstuck et al 2006	Disc bulge	No disc bulge	Worst pain in the last 2 wks (VAS, 0 to 10) (3 mo)	0.05 (p = 0.76) [a,b]
Kleinstuck et al 2006	Disc bulge	No disc bulge	Worst pain in the last 2 wks (VAS, 0 to 10) (12 mo)	0.13 (p = 0.42) [a,b]
Kleinstuck et al 2006	Disc bulge	No disc bulge	RMDQ (3 mo)	0.21 (p = 0.10) [a,b]
Kleinstuck et al 2006	Disc bulge	No disc bulge	RMDQ (12 mo)	0.22 (p = 0.11) [a,b]
de Schepper et al 2016	Disc bulge	No disc bulge	Recovery (GPE) (1 y)	1.3 (0.9 to 1.8) (OR)
Kleinstuck et al 2006	Disc degeneration	No disc degeneration	Average pain in the past 2 wks (VAS, 0 to 10) (3 mo)	0.10 (p = 0.57) [a,b]
Kleinstuck et al 2006	Disc degeneration	No disc degeneration	Worst pain in the last 2 wks (VAS, 0 to 10) (3 mo)	0.17 (p = 0.37) [a,b]
Kleinstuck et al 2006	Disc degeneration	No disc degeneration	Average pain in the past 2 wks (VAS, 0 to 10) (12 mo)	–0.18 (p = 0.25) [a,b]
Kleinstuck et al 2006	Disc degeneration	No disc degeneration	Worst pain in the last 2 wks (VAS, 0 to 10) (12 mo)	–0.13 (p = 0.49) [a,b]
Jensen et al 2015	Disc degeneration (severe)	No disc degeneration	Change in NRS (14 mo)	0.0 (–1.5 to 1.5)[a,c]

Jensen et al 2014	Disc degeneration	No disc degeneration	Pain (LBP rating scale) (1 y)	8.1 (–1.2 to 18.1)
McNee et al 2011	Disc degeneration	No disc degeneration	LBP in past 4 wks (18.5 mo)	1.6 (0.9 to 2.8) (OR)
McNee et al 2011	Disc degeneration	No disc degeneration	LBP on > 14 d in past 4 wks (18.5 mo)	1.7 (1.0 to 2.9) (OR)
McNee et al 2011	Disc degeneration	No disc degeneration	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	2.3 (1.3 to 4.0) (OR)
Kasch et al 2021a	Disc degeneration	No disc degeneration	Pain (NRS) (6 y)	4.7 (0.3 to 9.1)
Kleinstuck et al 2006	Disc degeneration	No disc degeneration	RMDQ (3 mo)	0.05 ($p = 0.73$)[a,b]
Kleinstuck et al 2006	Disc degeneration	No disc degeneration	RMDQ (12 mo)	–0.24 ($p = 0.15$)[a,b]
Jensen et al 2014	Disc degeneration	No disc degeneration	Disability (RMDQ) (1 y)	10.4 (–9.6 to 29.7)
Cai et al 2018	Disc degeneration (severe)	Disc degeneration (mild)	Change in VAS (LBP) (6 mo)	–8.5 (–23.4 to 6.4)
Udby et al 2021	Disc degeneration	No disc degeneration	Disability (RMDQ) (13 y)	0.06 ($p = 0.41$)[d]
Suri et al 2016	Disc extrusion	No disc extrusion	LBP recurrence (4 y)	1.0 (0.6 to 1.6) (HR)
Suri et al 2016	Disc extrusion	No disc extrusion	Leg pain recurrence (4 y)	0.7 (0.5 to 1.2) (HR)
Kerr et al 2015	Disc extrusion	Disc protrusion	change in ODI (8 y)	–2.4 (–5.2 to 0.4)[e]
Kerr et al 2015	Disc extrusion	Disc sequestration	change in ODI (8 y)	1.2 (–3.6 to 6.0)[e]
Jensen et al 2014	Disc height reduction	No disc height reduction	Pain (LBP rating scale) (1 y)	4.3 (–4.7 to 13.4)
Kasch et al 2021a	Disc height loss	No disc height loss	Pain (NRS) (6 y)	–4.8 (–10.2 to 0.5)
Jensen et al 2014	Disc height reduction	No disc height reduction	Disability (RMDQ) (1 y)	5.3 (–11.6 to 22.2)
Hellum et al 2012	Disc height reduction (rehab group)	No disc height reduction	ODI (2 y)	1.4 (0.5 to 4.2) (OR)

Hellum et al 2012	Disc height reduction (surgery group)	No disc height reduction	ODI (2 y)	1.0 (0.4 to 2.9) (OR)
Jensen et al 2015	Disc herniation (large)	No large disc herniation	Change in NRS (14 mo)	0.7 (–0.8 to 2.2)[c,d]
Modic et al 2005	Disc herniation	No disc herniation	RMDQ (6 wks)	2.7 (1.4 to 5.1) (OR)[f]
de Schepper et al 2016	Disc herniation	No disc herniation	Recovery (GPE) (1 y)	0.5 (0.3 to 0.7) (OR)
de Schepper et al 2016	Disc herniation	No disc herniation	Recovery (GPE) (1 y)	1.6 (1.1 to 2.6) (OR)[g]
Kesikburun et al 2019	Disc herniation (non-regression)	Disc herniation (complete regression)	NRS (17 mo)	18.0 (–2.1 to 38.1)
Kesikburun et al 2019	Disc herniation (non-regression)	Disc herniation (partial regression)	NRS (17 mo)	11.0 (–12.1 to 34.1)
Kesikburun et al 2019	Disc herniation (partial regression)	Disc herniation (complete regression)	NRS (17 mo)	7.0 (–10.8 to 24.8)
Jensen et al 2014	Disc herniation (protrusion, extrusion, sequestration)	No disc herniation	Pain (LBP rating scale) (1 y)	11.2 (–0.7 to 23.0)
Kasch et al 2021a	Disc herniation	No disc herniation	Pain (NRS) (6 y)	–7.0 (–16.3 to 2.4)
Jensen et al 2014	Disc herniation (protrusion, extrusion, sequestration)	No disc herniation	Disability (RMDQ) (1 y)	22.0 (–0.2 to 44.2)
Kesikburun et al 2019	Disc herniation (non-regression)	Disc herniation (complete regression)	ODI (17 mo)	14.2 (–4.9 to 33.3)
Kesikburun et al 2019	Disc herniation (non-regression)	Disc herniation (partial regression)	ODI (17 mo)	9.8 (–13.4 to 33.0)
Kesikburun et al 2019	Disc herniation (partial regression)	Disc herniation (complete regression)	ODI (17 mo)	4.4 (–12.1 to 20.9)
Suri et al 2016	Disc herniation (posterolateral)	Other disc herniation	LBP recurrence (4 y)	0.7 (0.4 to 1.1) (HR)

Suri et al 2016	Disc herniation (posterolateral)	Other disc herniation	Leg pain recurrence (4 y)	0.6 (0.4 to 0.9) (HR)
Suri et al 2016	Disc herniation (posterolateral)	Other disc herniation	Leg pain recurrence (4 y)	0.6 (0.4 to 1.0) (HR)[h]
Kerr et al 2015	Disc herniation (posterolateral)	Disc herniation (other herniation location)	change in ODI (8 y)	–3.7 (–6.4 to –1.0)[e]
Kuligowski et al 2020	Disc protrusion	Disc extrusion	NRS (1 mo)	30.00 (not reported)
Kuligowski et al 2020	Disc protrusion	Disc extrusion	NRS (1 mo)	1.80 ($p = 0.06$)[i]
Kuligowski et al 2020	Disc protrusion	Disc extrusion	ODI (1 mo)	2.50 ($p = 0.01$)[i]
Kuligowski et al 2020	Disc protrusion	Disc extrusion	ODI (1 mo)	18.00 (not reported)
Kerr et al 2015	Disc sequestration	Disc protrusion	change in ODI (8 y)	–3.6 (–8.4 to 1.2)[e]
Suri et al 2016	Disc sequestration	No disc sequestration	LBP recurrence (4 y)	0.6 (0.2 to 1.4) (HR)
Suri et al 2016	Disc sequestration	No disc sequestration	Leg pain recurrence (4 y)	0.9 (0.4 to 1.9) (HR)
Hellum et al 2012	Facet joint arthropathy grade 2/3 (rehab group)	No facet joint arthropathy grade 2/3	ODI (2 y)	1.3 (0.3 to 5.3) (OR)
Hellum et al 2012	Facet joint arthropathy grade 2/3 (surgery group)	No facet joint arthropathy grade 2/3	ODI (2 y)	0.9 (0.2 to 3.9) (OR)
Udby et al 2021	Facet joint degeneration	No facet joint degeneration	Disability (RMDQ) (13 y)	–0.02 ($p = 0.80$)[j]
Jensen et al 2014	Osteophytes	No osteophytes	Pain (LBP rating scale) (1 y)	–1.8 (–9.5 to 5.9)
Jensen et al 2014	Osteophytes	No osteophytes	Disability (RMDQ) (1 y)	–0.5 (–14.8 to 13.8)
Kasch et al 2021a	Hypertrophy ligamentus flava	No hypertrophy ligamentus flava	Pain (NRS) (6 y)	8.4 (–3.7 to 20.4)
Kleinstuck et al. 2006	High-intensity zone	No high-intensity zone	Average pain in the past 2 wk (VAS, 0 to 10) (3 mo)	–0.14 ($p = 0.35$)[a,b]

Kleinstuck et al. 2006	High-intensity zone	No high-intensity zone	Average pain in the past 2 wks (VAS, 0 to 10) (12 mo)	–0.38 ($p = 0.01$)[a,b]
Kleinstuck et al. 2006	High-intensity zone	No high-intensity zone	Worst pain in the last 2 wks (VAS, 0 to 10) (3 mo)	–0.04 ($p = 0.80$)[a,b]
Kleinstuck et al. 2006	High-intensity zone	No high-intensity zone	Worst pain in the last 2 wks (VAS, 0 to 10) (12 mo)	–0.21 ($p = 0.19$)[a,b]
Jensen et al 2014	High-intensity zone	No high-intensity zone	Pain (LBP rating scale) (1 y)	4.2 (–4.2 to 12.6)
Kasch et al 2021a	High-intensity zone	No high-intensity zone	Pain (NRS) (6 y)	–0.3 (–4.3 to 3.8)
Jensen et al 2014	High-intensity zone	No high-intensity zone	Disability (RMDQ) (1 y)	5.0 (–9.5 to 19.4)
Kleinstuck et al 2006	High-intensity zone	No high-intensity zone	RMDQ (3 mo)	–0.20 ($p = 0.12$)[a,b]
Kleinstuck et al 2006	High-intensity zone	No high-intensity zone	RMDQ (12 mo)	–0.08 ($p = 0.59$)[a,b]
Wilkens et al 2013	High-intensity zone	No high-intensity zone	Disability (RMDQ) (1 y)	–1.1 (?)
Wilkens et al 2013	High-intensity zone	No high-intensity zone	Disability (RMDQ) (1 y)	–3.3 (–8.1 to 1.5)[k]
Hellum et al 2012	High-intensity zone (posterior) (rehab group)	No posterior high-intensity zone	ODI (2 y)	0.6 (0.2 to 1.8) (OR)
Hellum et al 2012	High-intensity zone (posterior) (surgery group)	No posterior high-intensity zone	ODI (2 y)	1.6 (0.7 to 3.7) (OR)
Annen et al 2016	Modic type 1 and 2	No Modic	NRS (LBP) (1 mo)	–3.0 (–16.6 to 10.6)
Peterson et al 2014	Modic type 1 and 2	No Modic	NRS (1 mo)	7.3 (1.1 to 13.5)
Annen et al 2018	Modic type 1 and 2	No Modic	Pain (NRS) (1 mo)	–1.8 (–14.9 to 11.3)

Annen et al 2018	Modic type 1 and 2	No Modic	Pain (NRS) (3 mo)	1.3 (–12.2 to 14.8)
Annen et al 2016	Modic type 1 and 2	No Modic	NRS (LBP) (3 mo)	0.7 (–13.2 to 14.6)
Chen et al 2019	Modic type 1 and 2	No Modic	VAS (3 mo)	8.0 (4.5 to 11.5)
Chen et al 2019	Modic type 1 and 2	No Modic	VAS (6 mo)	6.3 (2.8 to 9.8)
Annen et al 2016	Modic type 1 and 2	No Modic	NRS (LBP) (6 mo)	–2.0 (–15.3 to 11.3)
Annen et al 2016	Modic type 1 and 2	No Modic	NRS (LBP) (1 y)	–3.7 (–19.5 to 12.1)
Jensen et al 2014	Modic type 1 and 2	No Modic	Pain (LBP rating scale) (1 y)	–18.0 (–29.2 to –6.8)
Jensen et al 2014	Modic type 1 and 2	No Modic	Pain (LBP rating scale) (1 y)	–17.3 (–28.6 to –5.9)[l]
Jensen et al 2014	Modic type 1 and 2	No Modic	Pain (LBP rating scale) (1 y)	–12.1 (–22.8 to –1.4)[m]
el Barzouhi et al 2014	Modic type 1 and 2	No Modic	Disabling LBP ≥ 40/100 mm (1 y)	1.4 (0.3 to 6.3) (OR)
Romero-Muñoz et al 2017	Modic type 1 and 2	No Modic	VAS (LBP) (10 y)	–1.0 (–14.3 to 12.3)
Chen et al 2019	Modic type 1 and 2	No Modic	VAS (Improvement rate % from baseline to 3 mo)	–12.5%
Chen et al 2019	Modic type 1 and 2	No Modic	VAS (Improvement rate % from 3 to 6 mo)	4.8%
Chen et al 2019	Modic type 1 and 2	No Modic	VAS (Improvement rate % from baseline to 6 mo)	–9.7%
Annen et al 2016	Modic type 1 and 2	No Modic	NRS (Leg pain) (1 mo)	4.6 (–10.0 to 19.2)

Annen et al 2016	Modic type 1 and 2	No Modic	NRS (Leg pain) (3 mo)	2.7 (–12.9 to 18.3)
Annen et al 2016	Modic type 1 and 2	No Modic	NRS (Leg pain) (6 mo)	10.0 (–4.5 to 24.5)
Annen et al 2016	Modic type 1 and 2	No Modic	NRS (Leg pain) (1 y)	7.1 (–8.0 to 22.2)
Romero-Muñoz et al 2017	Modic type 1 and 2	No Modic	VAS (Leg pain) (10 y)	0.0 (–15.4 to 15.4)
Annen et al 2016	Modic type 1 and 2	No Modic	ODI (1 mo)	8.4 (–0.6 to 17.3)
Annen et al 2016	Modic type 1 and 2	No Modic	ODI (3 mo)	10.1 (0.3 to 20.0)
Chen et al 2019	Modic type 1 and 2	No Modic	ODI (3 mo)	3.8 (2.0 to 5.7)
Chen et al 2019	Modic type 1 and 2	No Modic	ODI (6 mo)	2.9 (0.9 to 4.8)
Annen et al 2016	Modic type 1 and 2	No Modic	ODI (6 mo)	15.3 (6.3 to 24.3)
Annen et al 2016	Modic type 1 and 2	No Modic	ODI (1 y)	7.3 (–3.5 to 18.0)
Jensen et al 2014	Modic type 1 and 2	No Modic	Disability (RMDQ) (1 y)	–45.4 (–66.7 to 24.1)
Jensen et al 2014	Modic type 1 and 2	No Modic	Disability (RMDQ) (1 y)	–45.8 (–67.1 to –24.5)[l]
Jensen et al 2014	Modic type 1 and 2	No Modic	Disability (RMDQ) (1 y)	–39.1 (–60.8 to –17.5)[m]
Hellum et al 2012	Modic type 1 and 2 (rehab group)	No Modic	ODI (2 y)	1.0 (0.3 to 3.5) (OR)
Hellum et al 2012	Modic type 1 and 2 (surgery group)	No Modic	ODI (2 y)	1.1 (0.3 to 4.0) (OR)

Romero-Muñoz 2017	Modic type 1 and 2	No Modic	ODI (10 y)	–13.2 (–26.9 to 0.5)
Chen et al 2019	Modic type 1 and 2	No Modic	ODI (Improvement rate % from baseline to 3 mo)	–9.0%
Chen et al 2019	Modic type 1 and 2	No Modic	ODI (Improvement rate % from 3 to 6 mo)	4.4%
Chen et al 2019	Modic type 1 and 2	No Modic	ODI (Improvement rate % from baseline to 6 mo)	–6.8%
Peterson et al 2014	Modic type 1 and 2	No Modic	PGIG (1 mo)	5.1 (0.2 to 10.1)
Annen et al 2018	Modic type 1 and 2	No Modic	Improvement (PGIC) (1 mo)	1.1 (0.4 to 2.7) (OR)
Annen et al 2018	Modic type 1 and 2	No Modic	Improvement (PGIC) (3 mo)	1.3 (0.5 to 3.3) (OR)
Annen et al 2016	Modic type 1 and 2	No Modic	Recovery (PGIC) (1 mo)	0.8 (0.3 to 2.3) (OR)
Annen et al 2016	Modic type 1 and 2	No Modic	Recovery (PGIC) (3 mo)	1.6 (0.4 to 6.8) (OR)
Annen et al 2016	Modic type 1 and 2	No Modic	Recovery (PGIC) (6 mo)	1.7 (0.3 to 9.5) (OR)
Annen et al 2016	Modic type 1 and 2	No Modic	Recovery (PGIC) (1 y)	6.3 (0.7 to 54.6) (OR)
Peterson et al 2014	Modic type 1 and 2	No Modic	Recovery (PGIC) (1 mo)	1.6 (1.0 to 2.4) (OR)
Annen et al 2018	Modic type 1 and 2	No Modic	Bournemouth Questionnaire (1 mo)	–5.4 (–12.4 to 1.6)
Annen et al 2018	Modic type 1 and 2	No Modic	Bournemouth Questionnaire (3 mo)	–3.9 (–10.9 to 3.1)
Kleinstuck et al 2006	Modic type 1 and/or 2	No Modic	Average pain in the past 2 wks (VAS, 0 to 10) (3 mo)	–0.25 ($p = 0.16$)[a,b]
Kleinstuck et al 2006	Modic type 1 and/or 2	No Modic	Average pain in the past 2 wks (VAS, 0 to 10) (12 mo)	0.17 ($p = 0.26$)[a,b]

Kleinstuck et al 2006	Modic type 1 and/or 2	No Modic	Worst pain in the last 2 wks (VAS, 0 to 10) (3 mo)	–0.27 ($p = 0.15$)[a,b]
Kleinstuck et al 2006	Modic type 1 and/or 2	No Modic	Worst pain in the last 2 wks (VAS, 0 to 10) (12 mo)	0.20 ($p = 0.28$)[a,b]
Kleinstuck et al 2006	Modic type 1 and/or 2	No Modic	RMDQ (3 mo)	0.10 ($p = 0.50$)[a,b]
Kleinstuck et al 2006	Modic type 1 and/or 2	No Modic	RMDQ (12 mo)	0.20 ($p = 0.23$)[a,b]
Jensen et al 2014	Modic changes (type 1, or type 1 and 2)	No Modic	Pain (LBP rating scale) (1 y)	–9.8 (–16.7 to –2.9)
Jensen et al 2014	Modic changes (type 1, or type 1 and 2)	No Modic	Disability (RMDQ) (1 y)	–14.6 (–27.8 to –1.4)
Hellum et al 2012	Modic type 1 or 2 (rehab group)	No Modic	ODI (2 y)	3.4 (0.6 to 17.7) (OR)
Hellum et al 2012	Modic type 1 or 2 (surgery group)	No Modic	ODI (2 y)	5.3 (1.1 to 25.3) (OR)
Wilkens et al 2013	Modic type 1, 2, or 3	No Modic	Disability (RMDQ) (1 y)	3.8 (–3.0 to 10.7)
Wilkens et al 2013	Modic type 1, 2, or 3	No Modic	Disability (RMDQ) (1 y)	1.9 (–4.9 to 8.7)[k]
Annen et al 2016	Modic type 1	No Modic	NRS (LBP) (1 mo)	–8.5 (–26.5 to 9.5)
Annen et al 2018	Modic type 1	No Modic	Pain (NRS) (1 mo)	–7.4 (–21.1 to 6.3)
Schistad et al 2014	Modic type 1	No Modic	Pain (VAS) (6 wks)	6.0 (–6.3 to 18.3)
Annen et al 2016	Modic type 1	No Modic	NRS (LBP) (3 mo)	–4.0 (–22.6 to 14.6)
Annen et al 2018	Modic type 1	No Modic	Pain (NRS) (3 mo)	–1.6 (–16.2 to 13.0)
Chen et al 2019	Modic type 1	No Modic	VAS (3 mo)	5.0 (1.4 to 8.6)

Chen et al 2019	Modic type 1	No Modic	VAS (6 mo)	2.0 (–2.4 to 6.4)
Schistad et al 2014	Modic type 1	No Modic	Pain (VAS) (6 mo)	7.0 (–5.3 to 19.3)
Annen et al 2016	Modic type 1	No Modic	NRS (LBP) (6 mo)	–1.5 (–19.3 to 16.3)
Cai et al 2018	Modic type 1	No Modic	Change in VAS (LBP) (6 mo)	0.3 (–15.7 to 16.3)
Annen et al 2016	Modic type 1	No Modic	NRS (LBP) (1 y)	–4.5 (–25.9 to 16.9)
Schistad et al 2014	Modic type 1	No Modic	Pain (VAS) (1 y)	1.0 (–12.0 to 14.0)
Keller at al 2012	Modic type 1	No Modic	Pain at rest during previous wk (NPRS, 0 to 10) (12 mo)	2.0 (–6.2 to 10.3)
Keller at al 2012	Modic type 1	No Modic	Pain at activity during previous wk (NPRS, 0 to 10) (12 mo)	1.0 (0.8 to 1.2) (HR)
Keller at al 2012	Modic type 1	No Modic	Pain at rest during previous wk (NPRS, 0 to 10) (12 mo)	0.9 (0.8 to 1.1) (HR)
Jensen et al 2014	Modic type 1	No Modic	Pain (LBP rating scale) (1 y)	–14.5 (–27.2 to –1.8)
Jensen et al 2014	Modic type 1	No Modic	Pain (LBP rating scale) (1 y)	–14.2 (–27.0 to –1.3)[l]
Jensen et al 2014	Modic type 1	No Modic	Pain (LBP rating scale) (1 y)	–11.0 (–23.7 to 1.7)[m]
Jensen et al 2015	Modic type 1	No Modic	Change in NRS (14 mo)	–1.8 (–3.4 to –0.2)[c]
Jensen et al 2015	Modic type 1 (exercise group)	No Modic	Change in NRS (14 mo)	17.8 (2.6 to 33.0)
Jensen et al 2015	Modic type 1 (large ≥25 (rest + exercise group)	No Modic	Change in NRS (14 mo)	7.0 (–2.7 to 16.7)
Jensen et al 2015	Modic type 1 (large) (exercise group)	No Modic	Change in NRS (14 mo)	–0.7 (–14.6 to 13.2)

Jensen et al 2015	Modic type 1 (large) (rest group)	No Modic	Change in NRS (14 mo)	14.3 (0.8 to 27.9)
Jensen et al 2015	Modic Type 1 (rest + exercise group)	No Modic	Change in NRS (14 mo)	22.8 (12.2 to 33.4)
Jensen et al 2015	Modic type 1 (rest group)	No Modic	Change in NRS (14 mo)	27.8 (12.8 to 42.8)
Jensen et al 2012	Modic type 1	No Modic	Worsening of LBP intensity (NPRS, 0 to 10) (14 mo)	6.2 (1.9 to 20.3) (OR)
Jensen et al 2012	Modic type 1	No Modic	Change in LBP intensity (NPRS, 0 to 10) (14 mo)	1.0 (0.4 to 2.4) (OR)
el Barzouhi et al 2014	Modic type 1	No Modic	Disabling LBP ≥ 40/100 mm (1 y)	1.5 (0.1 to 16.0) (OR)
Kasch et al 2021a	Modic type 1	No Modic	Pain (NRS) (6 y)	–3.0 (–13.4 to 7.4)
Romero-Muñoz et al 2017	Modic type 1	No Modic	VAS (LBP) (10 y)	–5.0 (–23.6 to 13.6)
Annen et al 2016	Modic type 1	No Modic	NRS (Leg pain) (1 mo)	–6.4 (–24.8 to 12.0)
Annen et al 2016	Modic type 1	No Modic	NRS (Leg pain) (1 y)	–4.8 (–24.8 to 15.2)
Annen et al 2016	Modic type 1	No Modic	NRS (Leg pain) (3 mo)	–8.0 (–27.9 to 11.9)
Annen et al 2016	Modic type 1	No Modic	NRS (Leg pain) (6 mo)	4.1 (–14.8 to 23.0)
Romero-Muñoz et al 2017	Modic type 1	No Modic	VAS (Leg pain) (10 y)	11.0 (–9.7 to 31.7)
Chen et al 2019	Modic type 1	No Modic	VAS (Improvement rate % from baseline to 3 mo)	–7.2%
Chen et al 2019	Modic type 1	No Modic	VAS (Improvement rate % from 3 to 6 mo)	9.6%

Chen et al 2019	Modic type 1	No Modic	VAS (Improvement rate % from baseline to 6 mo)	–2.6%
Annen et al 2016	Modic type 1	No Modic	ODI (1 mo)	5.7 (–5.8 to 17.2)
Annen et al 2016	Modic type 1	No Modic	ODI (3 mo)	9.3 (–2.2 to 20.9)
Chen et al 2019	Modic type 1	No Modic	ODI (3 mo)	0.8 (–1.2 to 2.8)
Chen et al 2019	Modic type 1	No Modic	ODI (6 mo)	–1.4 (–3.5 to 0.7)
Annen et al 2016	Modic type 1	No Modic	ODI (6 mo)	15.5 (5.3 to 25.7)
Annen et al 2016	Modic type 1	No Modic	ODI (1 y)	2.9 (–10.6 to 16.5)
Wilkens et al 2013	Modic type 1	No Modic	Disability (RMDQ) (1 y)	–5.2 (–13.8 to 3.4)
Wilkens et al 2013	Modic type 1	No Modic	Disability (RMDQ) (1 y)	–4.5 (–12.8 to 3.9)[k]
Jensen et al 2014	Modic type 1	No Modic	Disability (RMDQ) (1 y)	–22.5 (–47.1 to 2.1)
Jensen et al 2014	Modic type 1	No Modic	Disability (RMDQ) (1 y)	–21.0 (–45.8 to 3.8)[l]
Jensen et al 2014	Modic type 1	No Modic	Disability (RMDQ) (1 y)	–19.5 (–46.3 to 7.1)[m]
Romero-Muñoz et al 2017	Modic type 1	No Modic	ODI (10 y)	–8.6 (–26.0 to 8.8)
Keller at al 2012	Modic type 1	No Modic	ODI (12 mo)	0.6 (–4.1 to 5.2)
Keller at al 2012	Modic type 1	No Modic	Disability (ODI Norwegian version, 0 to 100) (12 mo)	1.0 (1.0 to 1.0) (HR)

Hellum et al 2012	Modic type 1 (rehab group)	No Modic	ODI (2 y)	0.7 (0.2 to 1.9) (OR)
Hellum et al 2012	Modic type 1 (surgery group)	No Modic	ODI (2 y)	2.0 (0.7 to 6.2) (OR)
Chen et al 2019	Modic type 1	No Modic	ODI (Improvement rate % from baseline to 3 mo)	–3.1%
Chen et al 2019	Modic type 1	No Modic	ODI (Improvement rate % from 3 to 6 mo)	–11.6%
Chen et al 2019	Modic type 1	No Modic	ODI (Improvement rate % from baseline to 6 mo)	1.9%
Udby et al 2021	Modic change (1,2,3)	No Modic change	Disability (RMDQ) (13 y)	0.31 (not reported) (R^2)
Udby et al 2021	Modic change (1,2,3)	No Modic change	Disability (RMDQ) (13 y)	0.15 ($p = 0.03$)[d]
Annen et al 2016	Modic type 1	No Modic	Recovery (PGIC) (1 mo)	2.1 (0.6 to 7.3) (OR)
Annen et al 2018	Modic type 1	No Modic	Improvement (PGIC) (1 mo)	1.1 (0.4 to 3.4) (OR)
Annen et al 2018	Modic type 1	No Modic	Improvement (PGIC) (3 mo)	1.3 (0.4 to 3.8) (OR)
Annen et al 2016	Modic type 1	No Modic	Recovery (PGIC) (3 mo)	2.1 (0.4 to 11.9) (OR)
Annen et al 2016	Modic type 1	No Modic	Recovery (PGIC) (6 mo)	1.6 (0.2 to 13.0) (OR)
Annen et al 2016	Modic type 1	No Modic	Recovery (PGIC) (1 y)	20.3 (2.1 to 193.9) (OR)
Annen et al 2018	Modic type 1	No Modic	Bournemouth Questionnaire (1 mo)	–8.2 (–16.9 to 0.5)
Annen et al 2018	Modic type 1	No Modic	Bournemouth Questionnaire (3 mo)	–3.0 (–11.3 to 5.3)
Keller at al 2012	Modic type 1	No Modic	Education (University education) (12 mo)	2.2 (1.0 to 4.7) (HR)

Keller at al 2012	Modic type 1	No Modic	FABQ-PA (12 mo)	1.0 (1.0 to 1.1) (HR)
Keller at al 2012	Modic type 1	No Modic	FABQ-work (12 mo)	1.0 (0.9 to 1.0) (HR)
Jensen et al 2015	Modic (large) (exercise group)	Modic (small) (exercise group)	Change in NRS (14 mo)	–12.3 (–27.4 to 2.8)
Jensen et al 2015	Modic (large) (rest group)	Modic (small) (rest group)	Change in NRS (14 mo)	2.5 (–12.2 to 17.2)
Jensen et al 2015	Large Modic change (any type of MC)	Small Modic change (any type of MC)	Change in NRS (14 mo)	1.2 (–0.4 to 2.8)[b,c]
Jensen et al 2015	Large Modic Type 1	No large Modic Type 1	Change in NRS (14 mo)	0.1 (–1.4 to 1.6)[b,c]
Jensen et al 2015	Modic (large ≥ 25% of vertebral height) (rest + exercise group)	Modic (small < 25% of vertebral height) (rest + exercise group)	Change in NRS (14 mo)	–4.7 (–15.2 to 5.8)
Annen et al 2016	Modic type 1	Modic type 2	NRS (LBP) (1 mo)	–9.4 (–33.8 to 15.0)
Annen et al 2018	Modic type 1	Modic type 2	Pain (NRS) (1 mo)	–11.1 (–30.1 to 7.9)
Annen et al 2018	Modic type 1	Modic type 2	Pain (NRS) (3 mo)	–5.7 (–22.9 to 11.5)
Chen et al 2019	Modic type 1	Modic type 2	VAS (3 mo)	–5.0 (–9.8 to –0.2)
Annen et al 2016	Modic type 1	Modic type 2	NRS (LBP) (3 mo)	–7.7 (–30.0 to 14.6)
Chen et al 2019	Modic type 1	Modic type 2	VAS (6 mo)	–7.0 (–11.6 to –2.4)
Annen et al 2016	Modic type 1	Modic type 2	NRS (LBP) (6 mo)	0.8 (–19.6 to 21.2)
Annen et al 2016	Modic type 1	Modic type 2	NRS (LBP) (1 y)	–1.3 (–28.4 to 25.8)

el Barzouhi et al 2014	Modic type 1	Modic type 2	Disabling LBP ≥ 40/100 mm (1 y)	1.0 (0.1 to 10.1) (OR)
Romero-Muñoz et al 2017	Modic type 1	Modic type 2	VAS (LBP) (10 y)	–8.0 (–26.6 to 10.6)
Chen et al 2019	Modic type 1	Modic type 2	VAS (Improvement rate % from baseline to 3 mo)	8.7%
Chen et al 2019	Modic type 1	Modic type 2	VAS (Improvement rate % from 3 to 6 mo)	7.9%
Chen et al 2019	Modic type 1	Modic type 2	VAS (Improvement rate % from baseline to 6 mo)	11.7%
Annen et al 2016	Modic type 1	Modic type 2	NRS (Leg pain) (1 mo)	–18.1 (–43.3 to 7.1)
Annen et al 2016	Modic type 1	Modic type 2	NRS (Leg pain) (3 mo)	–18.4 (–43.9 to 7.1)
Annen et al 2016	Modic type 1	Modic type 2	NRS (Leg pain) (6 mo)	–10.1 (–33.9 to 13.7)
Annen et al 2016	Modic type 1	Modic type 2	NRS (Leg pain) (1 y)	–18.9 (–43.5 to 5.7)
Romero-Muñoz et al 2017	Modic type 1	Modic type 2	VAS (Leg pain) (10 y)	14.0 (–7.3 to 35.3)
Annen et al 2016	Modic type 1	Modic type 2	ODI (1 mo)	–4.4 (–21.1 to 12.4)
Annen et al 2016	Modic type 1	Modic type 2	ODI (3 mo)	–1.3 (–18.3 to 15.6)
Chen et al 2019	Modic type 1	Modic type 2	ODI (3 mo)	–5.0 (–7.5 to –2.6)
Annen et al 2016	Modic type 1	Modic type 2	ODI (6 mo)	0.3 (–16.7 to 17.4)
Chen et al 2019	Modic type 1	Modic type 2	ODI (6 mo)	–7.0 (–9.4 to –4.6)

Annen et al 2016	Modic type 1	Modic type 2	ODI (1 y)	–6.9 (–28.6 to 14.9)
Romero-Muñoz et al 2017	Modic type 1	Modic type 2	ODI (10 y)	4.0 (–19.0 to 27.0)
Chen et al 2019	Modic type 1	Modic type 2	ODI (Improvement rate % from baseline to 3 mo)	9.7%
Chen et al 2019	Modic type 1	Modic type 2	ODI (Improvement rate % from 3 to 6 mo)	11.9%
Chen et al 2019	Modic type 1	Modic type 2	ODI (Improvement rate % from baseline to 6 mo)	14.3%
Annen et al 2016	Modic type 1	Modic type 2	Recovery (PGIC) (1 mo)	2.1 (0.6 to 7.3) (OR)
Annen et al 2018	Modic type 1	Modic type 2	Improvement (PGIC) (1 mo)	1.1 (0.3 to 3.6) (OR)
Annen et al 2018	Modic type 1	Modic type 2	Improvement (PGIC) (3 mo)	0.9 (0.3 to 2.9) (OR)
Annen et al 2016	Modic type 1	Modic type 2	Recovery (PGIC) (3 mo)	1.2 (0.2 to 6.3) (OR)
Annen et al 2016	Modic type 1	Modic type 2	Recovery (PGIC) (6 mo)	1.0 (0.1 to 6.5) (OR)
Annen et al 2016	Modic type 1	Modic type 2	Recovery (PGIC) (1 y)	16.5 (1.7 to 159.1) (OR)
Annen et al 2018	Modic type 1	Modic type 2	Bournemouth Questionnaire (1 mo)	–5.3 (–14.9 to 4.4)
Annen et al 2018	Modic type 1	Modic type 2	Bournemouth Questionnaire (3 mo)	1.7 (–6.9 to 10.3)
Schistad 2014	Modic type 1	Modic type 2 and 3	Pain (VAS) (6 wk)	8.0 (–1.2 to 17.2)
Schistad 2014	Modic type 1	Modic type 2 and 3	Pain (VAS) (6 mo)	8.0 (–1.0 to 17.0)
Schistad 2014	Modic type 1	Modic type 2 and 3	Pain (VAS) (1 y)	0.0 (–10.5 to 10.5)

Romero-Muñoz et al 2017	Modic type 1	Modic type 3	VAS (LBP) (10 y)	18.0 (–47.4 to 83.4)
Romero-Muñoz et al 2017	Modic type 1	Modic type 3	VAS (Leg pain) (10 y)	40.0 (17.5 to 62.5)
Romero-Muñoz et al 2017	Modic type 1	Modic type 3	ODI (10 y)	38.0 (–40.0 to 116.0)
Annen et al 2016	Modic type 2	No Modic	NRS (LBP) (1 mo)	0.9 (–15.4 to 17.2)
Annen et al 2018	Modic type 2	No Modic	Pain (NRS) (1 mo)	3.7 (–12.6 to 20.0)
Annen et al 2018	Modic type 2	No Modic	Pain (NRS) (3 mo)	4.1 (–12.8 to 21.0)
Annen et al 2016	Modic type 2	No Modic	NRS (LBP) (3 mo)	3.7 (–14.7 to 22.1)
Chen et al 2019	Modic type 2	No Modic	VAS (3 mo)	10.0 (6.0 to 14.0)
Chen et al 2019	Modic type 2	No Modic	VAS (6 mo)	9.0 (5.6 to 12.4)
Annen et al 2016	Modic type 2	No Modic	NRS (LBP) (6 mo)	–2.3 (–20.5 to 15.9)
Annen et al 2016	Modic type 2	No Modic	NRS (LBP) (1 y)	–3.2 (–23.2 to 16.8)
Jensen et al 2014	Modic type 2	No Modic	Pain (LBP rating scale) (1 y)	–6.1 (–13.7 to 1.4)
Jensen et al 2014	Modic type 2	No Modic	Pain (LBP rating scale) (1 y)	–6.5 (–14.1 to 1.2)[l]
Jensen et al 2014	Modic type 2	No Modic	Pain (LBP rating scale) (1 y)	–1.1 (–8.3 to 6.1)[m]

el Barzouhi et al 2014	Modic type 2	No Modic	Disabling LBP ≥ 40/100 mm (1 y)	1.5 (0.3 to 7.1) (OR)
Keller at al 2012	Modic type 2	No Modic	Pain at rest during previous wk (NPRS, 0 to 10) (12 mo)	1.0 (–5.1 to 7.6)
Keller at al 2012	Modic type 2	No Modic	Pain at activity during previous wk (NPRS, 0 to 10) (12 mo)	0.9 (0.8 to 1.1) (HR)
Romero-Muñoz et al 2017	Modic type 2	No Modic	VAS (LBP) (10 y)	3.0 (–12.2 to 18.2)
Chen et al 2019	Modic type 2	No Modic	VAS (Improvement rate % from baseline to 3 mo)	–15.9%
Chen et al 2019	Modic type 2	No Modic	VAS (Improvement rate % from 3 to 6 mo)	1.7%
Chen et al 2019	Modic type 2	No Modic	VAS (Improvement rate % from baseline to 6 mo)	–14.3%
Annen et al 2016	Modic type 2	No Modic	NRS (Leg pain) (1 mo)	11.7 (–5.9 to 29.3)
Annen et al 2016	Modic type 2	No Modic	NRS (Leg pain) (3 mo)	10.4 (–9.4 to 30.2)
Annen et al 2016	Modic type 2	No Modic	NRS (Leg pain) (6 mo)	14.2 (–4.6 to 33.0)
Annen et al 2016	Modic type 2	No Modic	NRS (Leg pain) (1 y)	14.1 (–5.0 to 33.2)
Romero-Muñoz et al 2017	Modic type 2	No Modic	VAS (Leg pain) (10 y)	–3.0 (–18.9 to 12.9)
Annen et al 2016	Modic type 2	No Modic	ODI (1 mo)	10.1 (–0.2 to 20.3)
Annen et al 2016	Modic type 2	No Modic	ODI (3 mo)	10.7 (–1.8 to 23.1)
Chen et al 2019	Modic type 2	No Modic	ODI (3 mo)	5.8 (3.8 to 7.8)

Annen et al 2016	Modic type 2	No Modic	ODI (6 mo)	15.1 (4.7 to 25.6)
Chen et al 2019	Modic type 2	No Modic	ODI (6 mo)	5.6 (3.8 to 7.4)
Annen et al 2016	Modic type 2	No Modic	ODI (1 y)	9.8 (–1.4 to 21.0)
Keller at al. 2012	Modic type 2	No Modic	ODI (12 mo)	0.5 (–4.1 to 5.2)
Wilkens 2013	Modic type 2	No Modic	Disability (RMDQ) (1 y)	0.2 (–5.1 to 5.5)
Wilkens 2013	Modic type 2	No Modic	Disability (RMDQ) (1 y)	–2.3 (–7.4 to 2.8)[k]
Jensen et al 2014	Modic type 2	No Modic	Disability (RMDQ) (1 y)	–6.3 (–19.5 to 6.9)
Jensen et al 2014	Modic type 2	No Modic	Disability (RMDQ) (1 y)	–10.4 (–24.5 to 3.7)[l]
Jensen et al 2014	Modic type 2	No Modic	Disability (RMDQ) (1 y)	–5.5 (–19.7 to 8.5)[m]
Hellum et al 2012	Modic type 2 (rehab group)	No Modic	ODI (2 y)	2.1 (0.7 to 5.9) (OR)
Hellum et al 2012	Modic type 2 (surgery group)	No Modic	ODI (2 y)	1.3 (0.5 to 3.5) (OR)
Romero-Muñoz et al 2017	Modic type 2	No Modic	ODI (10 y)	–12.6 (–27.1 to 1.9)
Chen et al 2019	Modic type 2	No Modic	ODI (Improvement rate % from baseline to 3 mo)	–12.8%
Chen et al 2019	Modic type 2	No Modic	ODI (Improvement rate % from 3 to 6 mo)	–0.3%
Chen et al 2019	Modic type 2	No Modic	ODI (Improvement rate % from baseline to 6 mo)	–12.3%
Annen et al 2016	Modic type 2	No Modic	Recovery (PGIC) (1 mo)	1.0 (0.4 to 2.8) (OR)

Annen et al 2018	Modic type 2	No Modic	Improvement (PGIC) (1 mo)	1.0 (0.4 to 3.0) (OR)
Annen et al 2018	Modic type 2	No Modic	Improvement (PGIC) (3 mo)	1.4 (0.5 to 4.0) (OR)
Annen et al 2016	Modic type 2	No Modic	Recovery (PGIC) (3 mo)	1.8 (0.4 to 8.8) (OR)
Annen et al 2016	Modic type 2	No Modic	Recovery (PGIC) (6 mo)	1.7 (0.3 to 11.4) (OR)
Annen et al 2016	Modic type 2	No Modic	Recovery (PGIC) (1 y)	1.2 (0.1 to 20.8) (OR)
Keller at al 2012	Modic type 2	No Modic	Education (University education) (12 mo)	2.1 (1.0 to 4.3) (HR)
Keller at al 2012	Modic type 2	No Modic	FABQ-PA (12 mo)	1.0 (0.9 to 1.1) (HR)
Keller at al 2012	Modic type 2	No Modic	FABQ-work (12 mo)	1.0 (0.9 to 1.0) (HR)
Annen et al 2018	Modic type 2	No Modic	Bournemouth Questionnaire (1 mo)	–2.9 (–10.8 to 4.9)
Annen et al 2018	Modic type 2	No Modic	Bournemouth Questionnaire (3 mo)	–4.7 (–13.3 to 4.0)
Schistad et al 2014	Modic type 2 and 3	No Modic	Pain (VAS) (6 wks)	–2.0 (–9.0 to 5.0)
Schistad et al 2014	Modic type 2 and 3	No Modic	Pain (VAS) (6 mo)	–1.0 (–8.1 to 6.1)
Schistad et al 2014	Modic type 2 and 3	No Modic	Pain (VAS) (1 y)	1.0 (–7.1 to 9.1)
Kasch et al 2021a	Modic type 2 or 3	No Modic	Pain (NRS) (6 y)	1.2 (–4.1 to 6.4)
Romero-Muñoz et al 2017	Modic type 2	Modic type 3	VAS (LBP) (10 y)	26.0 (–64.2 to 116.2)

Romero-Muñoz et al 2017	Modic type 2	Modic type 3	VAS (Leg pain) (10 y)	26.0 (–5.1 to 57.1)
Romero-Muñoz et al 2017	Modic type 2	Modic type 3	ODI (10 y)	34.0 (–75.5 to 143.5)
Romero-Muñoz et al 2017	Modic type 3	No Modic	VAS (LBP) (10 y)	–23.0 (–63.9 to 17.9)
Romero-Muñoz et al 2017	Modic type 3	No Modic	VAS (Leg pain) (10 y)	–29.0 (–74.1 to 16.1)
Romero-Muñoz et al 2017	Modic type 3	No Modic	ODI (10 y)	–46.6 (–87.5 to –5.7)
Kovisto et al 2017	Modic type 1 (placebo group) (change in volume of M1)	Not applicable as change measure	Change in LBP intensity (11.9 mo)	0.28 (not reported)[j]
Jarvinen et al 2015	Modic type 1 (change in extent of M1)	Not applicable as change measure	NRS (2 y)	0.26 ($p = 0.04$)[a]
Jarvinen et al 2015	Modic type 1 (change in extent of M1)	Not applicable as change measure	NRS (2 y)	–0.19 ($p = 0.28$)[a,n]
Kovisto et al 2017	Modic type 1 (placebo group) (change in volume of M1)	Not applicable as change measure	ODI (11.9 mo)	0.32 (not reported)[j]
Jarvinen et al 2015	Modic type 1 (change in extent of M1)	Not applicable as change measure	ODI (2 y)	0.30 ($p = 0.02$)[a]
Jarvinen et al 2015	Modic type 1 (change in extent of M1)	Not applicable as change measure	ODI (2 y)	0.53 ($p = 0.00$)[a,n]
Kovisto et al 2017	Modic type 2 (placebo group) (change in volume of M2)	Not applicable as change measure	Change in LBP intensity (11.9 mo)	–0.16 (not reported)[j]

Jarvinen et al 2015	Modic type 2 (change in extent of M2)	Not applicable as change measure	NRS (2 y)	–0.24 ($p = 0.05$)[a]
Jarvinen et al 2015	Modic type 2 (change in extent of M2)	Not applicable as change measure	NRS (2 y)	–0.19 ($p = 0.16$)[a,n]
Kovisto et al 2017	Modic type 2 (placebo group) (change in volume of M2)	Not applicable as change measure	ODI (11.9 mo)	–0.33 (not reported)[o]
Jarvinen et al 2015	Modic type 2 (change in extent of M2)	Not applicable as change measure	ODI (2 y)	–0.13 ($p = 0.30$)[a]
Jarvinen et al 2015	Modic type 2 (change in extent of M2)	Not applicable as change measure	ODI (2 y)	–0.14 ($p = 0.32$)[a,n]
Harrison et al 2020	Nerve root compression	No nerve root compression	NRS (1 y)	–7.0 (–17.0 to 3.0)
Harrison et al 2020	Nerve root compression	No nerve root compression	NRS (3 y)	–3.0 (–13.0 to 7.0)
Harrison et al 2020	Nerve root compression	No nerve root compression	NRS (4 mo)	–3.0 (–13.0 to 7.0)
Jensen et al 2014	Nerve root impingement	No nerve root impingement	Pain (LBP rating scale) (1 y)	2.5 (–5.5 to 10.4)
Jensen et al 2014	Nerve root impingement	No nerve root impingement	Disability (RMDQ) (1 y)	2.9 (–11.1 to 16.8)
Modic et al 2005	Nerve root compression	No nerve root compression	RMDQ (6 wks)	1.3 (0.7 to 2.5) (OR)[f]
de Schepper et al 2016	Nerve root compression	No nerve root compression	Recovery (GPE) (1 y)	0.4 (0.3 to 0.6) (OR)
de Schepper et al 2016	Nerve root compression	No nerve root compression	Recovery (GPE) (1 y)	2.2 (1.4 to 3.5) (OR)[g]
de Schepper et al 2016	Nerve root compression	No nerve root compression	Recovery (GPE) (1 y)	2.2 (1.4 to 3.4) (OR)[p]
Konstantinou et al 2018	Nerve root compression	No nerve root compression (all patients)	RMDQ (4 mo)	0.9 (0.4 to 2.0) (OR)

Konstantinou et al 2018	Nerve root compression	No nerve root compression (all patients)	RMDQ (4 mo)	0.8 (0.4 to 1.7)[q] (OR)
Konstantinou et al 2018	Nerve root compression	No nerve root compression (sciatica)	RMDQ (4 mo)	1.0 (0.4 to 2.5) (OR)
Konstantinou et al 2018	Nerve root compression	No nerve root compression (sciatica)	RMDQ (4 mo)	0.9 (0.3 to 2.5)[q] (OR)
Konstantinou et al 2018	Nerve root compression	No nerve root compression (all patients)	RMDQ (12 mo)	0.9 (0.5 to 2.0) (OR)
Konstantinou et al 2018	Nerve root compression	No nerve root compression (all patients)	RMDQ (12 mo)	0.9 (0.5 to 2.0)[q] (OR)
Konstantinou et al 2018	Nerve root compression	No nerve root compression (sciatica)	RMDQ (12 mo)	0.9 (0.4 to 2.5) (OR)
Konstantinou et al 2018	Nerve root compression	No nerve root compression (sciatica)	RMDQ (12 mo)	1.1 (0.4 to 3.3)[q] (OR)
Hellum et al 2012	Nucleus pulposus grade 3/4 (rehab group)	No nucleus pulposus grade 3/4	ODI (2 y)	1.3 (0.3 to 5.3) (OR)
Hellum et al 2012	Nucleus pulposus grade 3/4 (surgery group)	No nucleus pulposus grade 3/4	ODI (2 y)	0.3 (0.0 to 2.4) (OR)
Jensen et al 2014	Spinal stenosis	No spinal stenosis	Pain (LBP rating scale) (1 y)	2.7 (–6.9 to 12.2)
Kasch et al 2021a	Spinal stenosis	No spinal canal stenosis	Pain (NRS) (6 y)	–2.4 (–7.7 to 2.9)
Jensen et al 2014	Spinal stenosis	No spinal stenosis	Disability (RMDQ) (1 y)	–6.0 (–23.5 to 11.5)
Modic et al 2005	Spinal stenosis (severe)	No spinal stenosis	RMDQ (6 wks)	0.4 (0.2 to 1.2) (OR)[f]
de Schepper et al 2016	Spinal stenosis	No spinal stenosis	Recovery (GPE) (1 y)	1.1 (0.7 to 2.0) (OR)
Kasch et al 2021a	Spondylolisthesis	No spondylolisthesis	Pain (NRS) (6 y)	–16.6 (–42.4 to 9.1)
de Schepper et al 2016	Spondylolisthesis	No spondylolisthesis	Recovery (GPE) (1 y)	1.4 (0.7 to 2.7) (OR)

Kasch et al 2021a	Schmorl lesion	No Schmorl lesion	Pain (NRS) (6 y)	–3.6 (–7.6 to 0.4)
McNee et al 2011	1 MRI abnormality	No MRI abnormality	LBP in past 4 wks (18.5 mo)	0.4 (0.1 to 1.2) (OR)
McNee et al 2011	1 MRI abnormality	No MRI abnormality	LBP on > 14 d in past 4 wks (18.5 mo)	0.6 (0.2 to 1.6) (OR)
McNee et al 2011	1 MRI abnormality	No MRI abnormality	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	1.0 (0.4 to 3.1) (OR)
Kasch et al 2021a	1 MRI finding	No MRI finding	Pain (NRS) (6 y)	–0.6 (–5.4 to 4.2)
McNee et al 2011	≥ 1 MRI abnormality	No MRI abnormality	LBP in past 4 wks (18.5 mo)	0.5 (0.2 to 1.3) (OR)
McNee et al 2011	≥ 1 MRI abnormality	No MRI abnormality	LBP on > 14 d in past 4 wks (18.5 mo)	0.8 (0.4 to 1.7) (OR)
McNee et al 2011	≥ 1 MRI abnormality	No MRI abnormality	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	1.3 (0.5 to 3.0) (OR)
McNee et al 2011	2 MRI abnormalities	No MRI abnormality	LBP in past 4 wks (18.5 mo)	0.5 (0.2 to 1.3) (OR)
McNee et al 2011	2 MRI abnormalities	No MRI abnormality	LBP on > 14 d in past 4 wks (18.5 mo)	0.6 (0.2 to 1.4) (OR)
McNee et al 2011	2 MRI abnormalities	No MRI abnormality	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	1.0 (0.4 to 2.7) (OR)
Kasch et al 2021a	2 MRI findings	No MRI finding	Pain (NRS) (6 y)	–1.9 (–7.8 to 4.0)
McNee et al 2011	≥ 2 MRI abnormalities	No MRI abnormality	LBP in past 4 wks (18.5 mo)	0.5 (0.2 to 1.4) (OR)
McNee et al 2011	≥ 2 MRI abnormalities	No MRI abnormality	LBP on > 14 d in past 4 wks (18.5 mo)	0.8 (0.4 to 1.8) (OR)
McNee et al 2011	≥ 2 MRI abnormalities	No MRI abnormality	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	1.3 (0.6 to 3.1) (OR)
McNee et al 2011	3 MRI abnormalities	No MRI abnormality	LBP in past 4 wks (18.5 mo)	0.5 (0.2 to 1.5) (OR)

McNee et al 2011	3 MRI abnormalities	No MRI abnormality	LBP on > 14 d in past 4 wks (18.5 mo)	1.0 (0.4 to 2.3) (OR)
McNee et al 2011	3 MRI abnormalities	No MRI abnormality	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	1.0 (0.4 to 2.7) (OR)
Kasch et al 2021a	3 MRI findings	No MRI finding	Pain (NRS) (6 y)	0.4 (–5.5 to 6.3)
McNee et al 2011	≥ 3 MRI abnormalities	No MRI abnormality	LBP in past 4 wks (18.5 mo)	0.6 (0.2 to 1.6) (OR)
McNee et al 2011	≥ 3 MRI abnormalities	No MRI abnormality	LBP on > 14 d in past 4 wks (18.5 mo)	1.0 (0.4 to 2.2) (OR)
McNee et al 2011	≥ 3 MRI abnormalities	No MRI abnormality	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	1.5 (0.6 to 3.8) (OR)
McNee et al 2011	4 MRI abnormalities	No MRI abnormality	LBP in past 4 wks (18.5 mo)	0.7 (0.2 to 2.1) (OR)
McNee et al 2011	4 MRI abnormalities	No MRI abnormality	LBP on > 14 d in past 4 wks (18.5 mo)	1.1 (0.4 to 2.7) (OR)
McNee et al 2011	4 MRI abnormalities	No MRI abnormality	Disabling LBP in past 4 wks (RMDQ > 11) (18.5 mo)	2.6 (1.0 to 7.1) (OR)
Kasch et al 2021a	4 MRI findings	No MRI finding	Pain (NRS) (6 y)	2.6 (–6.7 to 11.8)
Kasch et al 2021a	≥ 5 MRI findings	No MRI finding	Pain (NRS) (6 y)	–3.0 (–9.9 to 3.8)
		Populations with no LBP		
Hancock et al 2017	Disc annular tear	No disc annular tear	Days to recurrence of activity limiting LBP (1 y)	1.3 (0.4 to 3.6) (OR)
Hancock et al 2015	Disc annular tear	No disc annular tear	LBP recurrence (1 y)	1.0 (0.5 to 2.0) (HR)
Hancock et al 2017	Disc annular tear	No disc annular tear	Days to recurrence of activity limiting LBP (1 y)	1.2 (0.5 to 2.7) (HR)
Kasch et al 2021b	Disc degeneration	No disc degeneration	Pain (NRS) (6 y)	2.6 (–1.3 to 6.6)

Boos et al 2000	Disc degeneration	No disc degeneration	LBP (yes/no) (62 mo)	1.4 (0.4 to 4.6) (OR)
Boos et al 2000	Disc degeneration	No disc degeneration	LBP frequency (once/y) (62 mo)	0.5 (0.0 to 6.2) (OR)
Boos et al 2000	Disc degeneration	No disc degeneration	LBP frequency (<once/mo) (62 mo)	0.7 (0.2 to 2.9) (OR)
Boos et al 2000	Disc degeneration	No disc degeneration	LBP frequency (once/mo) (62 mo)	4.0 (0.2 to 96.8) (OR)
Boos et al 2000	Disc degeneration	No disc degeneration	LBP frequency (every wk) (62 mo)	2.0 (0.2 to 24.7) (OR)
Boos et al 2000	Disc degeneration	No disc degeneration	LBP duration (1 to 7 d) (62 mo)	0.4 (0.1 to 2.6) (OR)
Boos et al 2000	Disc degeneration	No disc degeneration	LBP duration (8 to 30 d) (62 mo)	2.2 (0.4 to 10.4) (OR)
Boos et al 2000	Disc degeneration	No disc degeneration	LBP duration (> 30 d) (62 mo)	2.2 (0.2 to 26.7) (OR)
Tonosu et al 2017	Disc degeneration	No disc degeneration	Back pain in the past 1 mo (10 ys)	1.3 (0.4 to 4.7) (OR)
Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP frequency: once/y (Nordic questionnaire) (60 mo)	2.4 (0.1 to 44.5) (OR)
Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP frequency: < once/mo (Nordic questionnaire) (60 mo)	3.6 (0.8 to 17.0) (OR)
Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP frequency: once/mo (Nordic questionnaire) (60 mo)	2.4 (0.1 to 44.5) (OR)
Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP frequency: every wk (Nordic questionnaire) (60 mo)	4.9 (0.4 to 62.6) (OR)
Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP duration : 1 to 7 d (Nordic questionnaire) (60 mo)	1.2 (0.2 to 8.2) (OR)
Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP duration : 8 to 30 d (Nordic questionnaire) (60 mo)	7.3 (1.2 to 45.3) (OR)
Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP duration : > 30 d (Nordic questionnaire) (60 mo)	4.9 (0.4 to 62.6) (OR)

Elfering et al 2002	Disc degeneration (worsening)	Disc degeneration (no change)	LBP (Nordic questionnaire) (60 mo)	2.9 (0.8 to 10.3) (OR)
Hancock et al 2015	Disc degeneration score Pfirrmann ≥ 3	Disc degeneration score Pfirrmann < 3	Recurrence of activity limiting LBP (1 y)	2.3 (0.3 to 18.3) (HR)[l]
Hancock et al 2017	Disc degeneration score Pfirrmann ≥ 3	Disc degeneration score Pfirrmann < 3	Days to recurrence of activity limiting LBP (1 y)	4.1 (0.5 to 35.9) (OR)
Hancock et al 2017	Disc degeneration score Pfirrmann ≥ 3	Disc degeneration score Pfirrmann < 3	Days to recurrence of activity limiting LBP (1 y)	3.6 (0.5 to 26.2) (HR)
Hancock et al 2015	Disc degeneration score Pfirrmann ≥ 3	Disc degeneration score Pfirrmann < 3	LBP recurrence (1 y)	2.6 (0.6 to 10.6) (HR)
Hancock et al 2015	Disc degeneration score Pfirrmann ≥ 3	Disc degeneration score Pfirrmann < 3	LBP recurrence (1 y)	1.9 (0.4 to 8.5) (HR)[r]
Kasch et al 2021b	Disc height loss	No disc height loss	Pain (NRS) (6 y)	1.9 (–3.6 to 7.4)
Hancock et al 2015	Disc height loss	No disc height loss	LBP recurrence (1 y)	1.3 (0.9 to 1.8) (HR)
Hancock et al 2017	Disc height loss	No disc height loss	Days to recurrence of activity limiting LBP (1 y)	3.3 (0.8 to 13.9) (HR)
Hancock et al 2017	Disc height loss	No disc height loss	Days to recurrence of activity limiting LBP (1 y)	4.1 (0.8 to 20.2) (OR)
Kasch et al 2021b	Disc herniation	No disc herniation	Pain (NRS) (6 y)	3.7 (–5.0 to 12.3)
Boos et al 2000	Disc herniation	No disc herniation	LBP (yes/no) (62 mo)	1.1 (0.3 to 4.2) (OR)
Boos et al 2000	Disc herniation	No disc herniation	LBP frequency (once/y) (62 mo)	0.7 (0.1 to 8.5) (OR)
Boos et al 2000	Disc herniation	No disc herniation	LBP frequency (< once/mo) (62 mo)	0.8 (0.2 to 3.9) (OR)
Boos et al 2000	Disc herniation	No disc herniation	LBP frequency (once/mo) (62 mo)	1.3 (0.1 to 33.2) (OR)
Boos et al 2000	Disc herniation	No disc herniation	LBP frequency (every wk) (62 mo)	0.7 (0.1 to 8.5) (OR)

Boos et al 2000	Disc herniation	No disc herniation	LBP duration (1 to 7 d) (62 mo)	0.9 (0.1 to 5.6) (OR)
Boos et al 2000	Disc herniation	No disc herniation	LBP duration (8 to 30 d) (62 mo)	1.2 (0.2 to 7.4) (OR)
Boos et al 2000	Disc herniation	No disc herniation	LBP duration (>30 d) (62 mo)	0.7 (0.1 to 9.0) (OR)
Hancock et al 2015	Disc herniation	No disc herniation	Recurrence of activity limiting LBP (1 y)	1.7 (0.4 to 8.0) (HR)[r]
Hancock et al 2015	Disc herniation	No disc herniation	LBP recurrence (1 y)	0.9 (0.4 to 2.3) (HR)
Hancock et al 2015	Disc herniation	No disc herniation	LBP recurrence (1 y)	0.9 (0.3 to 2.5) (HR)[r]
Hancock et al 2017	Disc herniation	No disc herniation	Days to recurrence of activity limiting LBP (1 y)	1.8 (0.4 to 7.5) (HR)
Hancock et al 2017	Disc herniation	No disc herniation	Days to recurrence of activity limiting LBP (1 y)	2.0 (0.4 to 10.6) (OR)
Kasch et al 2021b	Disc high-intensity zone	No high-intensity zone	Pain (NRS) (6 y)	3.4 (–0.3 to 7.0)
Hancock et al 2015	Disc high-intensity zone	No disc high-intensity zone	Recurrence of activity limiting LBP (1 y)	2.6 (1.1 to 5.7) (HR)[r]
Hancock et al 2015	Disc high-intensity zone	No disc high-intensity zone	LBP recurrence (1 y)	1.9 (1.0 to 3.5) (HR)
Hancock et al 2015	Disc high-intensity zone	No disc high-intensity zone	LBP recurrence (1 y)	1.8 (0.9 to 3.6) (HR)[r]
Hancock et al 2017	Disc high-intensity zone	No disc high-intensity zone	Days to recurrence of activity limiting LBP (1 y)	2.5 (1.2 to 5.1) (HR)
Hancock et al 2017	Disc high-intensity zone	No disc high-intensity zone	Days to recurrence of activity limiting LBP (1 y)	3.1 (1.2 to 8.3) (OR)
Jarvik et al 2005	Disc protrusion	No disc high-intensity zone	PFI (3 y)	0.5 (0.3 to 0.9) (HR)
Jarvik et al 2005	Disc extrusion	No disc extrusion	PFI (3 y)	1.2 (0.4 to 3.4) (HR)

Tonosu et al 2017	Disc bulge	No disc bulge	Back pain in the past 1 mo (10 ys)	0.6 (0.2 to 2.4) (OR)
Kasch et al 2021b	Hypertrophy ligamentus flava	No hypertrophy ligamentus flava	Pain (NRS) (6 y)	–4.1 (–17.3 to 9.1)
Hancock et al 2015	Facet joint arthrosis	No facet joint arthrosis	LBP recurrence (1 y)	1.0 (0.5 to 1.9) (HR)
Hancock et al 2015	Facet joint arthrosis	No facet joint arthrosis	LBP recurrence (1 y)	0.7 (0.3 to 1.6) (HR)[r]
Hancock et al 2015	Facet joint arthrosis	No facet joint arthrosis	Recurrence of activity limiting LBP (1 y)	0.4 (0.2 to 1.1) (HR)[r]
Hancock et al 2017	Facet joint osteoarthritis	No facet joint osteoarthritis	Days to recurrence of activity limiting LBP (1 y)	0.7 (0.3 to 1.7) (OR)
Hancock et al 2017	Facet joint osteoarthritis	No facet joint osteoarthritis	Days to recurrence of activity limiting LBP (1 y)	0.7 (0.3 to 1.5) (HR)
Kasch et al 2021b	Modic type 1	No Modic	Pain (NRS) (6 y)	1.2 (–8.4 to 10.8)
Hancock et al 2017	Modic type 1, 2, 3	No Modic	Days to recurrence of activity limiting LBP (1 y)	1.5 (0.5 to 5.0) (OR)
Hancock et al 2017	Modic type 1, 2, 3	No Modic	Days to recurrence of activity limiting LBP (1 y)	1.4 (0.6 to 3.4) (HR)
Hancock et al 2015	Modic type 1, 2, 3	No Modic	LBP recurrence (1 y)	1.1 (0.5 to 2.6) (HR)[r]
Hancock et al 2015	Modic type 1, 2, 3	No Modic	LBP recurrence (1 y)	1.0 (0.5 to 2.3) (HR)
Hancock et al 2015	Modic type 1, 2, 3	No Modic	Recurrence of activity limiting LBP (1 y)	1.4 (0.5 to 3.6) (HR)[r]
Kasch et al 2021b	Modic type 2 or 3	No Modic	Pain (NRS) (6 y)	8.1 (–1.7 to 17.8)
Boos et al 2000	Neural compromise	No neural compromise	LBP (yes/no) (62 mo)	2.3 (0.6 to 8.6) (OR)
Boos et al 2000	Neural compromise	No neural compromise	LBP frequency (once/y) (62 mo)	2.3 (0.2 to 30.6) (OR)

Boos et al 2000	Neural compromise	No neural compromise	LBP frequency (< once/mo) (62 mo)	2.0 (0.4 to 10.4) (OR)
Boos et al 2000	Neural compromise	No neural compromise	LBP frequency (once/mo) (62 mo)	4.6 (0.2 to 86.6) (OR)
Boos et al 2000	Neural compromise	No neural compromise	LBP frequency (every wk) (62 mo)	9.2 (0.7 to 122.4) (OR)
Boos et al 2000	Neural compromise	No neural compromise	LBP duration (1 to 7 d) (62 mo)	1.8 (0.3 to 11.8) (OR)
Boos et al 2000	Neural compromise	No neural compromise	LBP duration (8 to 30 d) (62 mo)	2.2 (0.4 to 11.9) (OR)
Boos et al 2000	Neural compromise	No neural compromise	LBP duration (> 30 d) (62 mo)	8.8 (0.7 to 117.2) (OR)
Jarvik et al 2005	Nerve root contact	No nerve root contact	PFI (3 y)	1.9 (0.6 to 5.8) (HR)
Kasch et al 2021b	Spinal canal stenosis	No spinal canal stenosis	Pain (NRS) (6 y)	7.0 (–0.2 to 14.2)
Hancock et al 2015	Spinal canal area	No spinal canal area	LBP recurrence (1 y)	0.7 (0.1 to 5.4) (HR)
Jarvik et al 2005	Spinal central stenosis	No spinal canal stenosis	PFI (3 y)	1.9 (0.8 to 4.8) (HR)
Kasch et al 2021b	Spondylolisthesis	No spondylolisthesis	Pain (NRS) (6 y)	11.1 (–0.6 to 22.8)
Hancock et al 2017	Spondylolisthesis	No spondylolisthesis	Days to recurrence of activity limiting LBP (1 y)	1.8 (0.6 to 5.9) (OR)
Hancock et al 2017	Spondylolisthesis	No spondylolisthesis	Days to recurrence of activity limiting LBP (1 y)	1.6 (0.7 to 3.6) (HR)
Hancock et al 2015	Spondylolisthesis	No spondylolisthesis	LBP recurrence (1 y)	1.2 (0.6 to 2.8) (HR)[r]
Hancock et al 2015	Spondylolisthesis	No spondylolisthesis	LBP recurrence (1 y)	1.3 (0.6 to 2.8) (HR)
Hancock et al 2015	Spondylolisthesis	No spondylolisthesis	Recurrence of activity limiting LBP (1 y)	1.8 (0.7 to 4.6) (HR)[r]

Kasch et al 2021b	Schmorl lesion	No Schmorl lesion	Pain (NRS) (6 y)	–5.3 (–8.6 to –1.9)
Borenstein et al 2001	Worsening abnormalities on MRI	No worsening abnormalities on MRI	LBP (7 y)	3.5 (not reported) (RR)
Kasch et al 2021b	1 MRI finding	No MRI finding	Pain (NRS) (6 y)	2.8 (–0.5 to 6.0)
Hancock et al 2017	1 finding present	0 findings	Days to recurrence of activity limiting LBP (1 y)	3.3 (0.4 to 30.3) (OR)
Hancock et al 2017	1 finding present	0 findings	Days to recurrence of activity limiting LBP (1 y)	2.8 (0.4 to 22.0) (HR)
Kasch et al 2021b	2 MRI findings	No MRI finding	Pain (NRS) (6 y)	2.1 (–2.1 to 6.4)
Hancock et al 2017	2 findings present	0 findings	Days to recurrence of activity limiting LBP (1 y)	8.6 (1.0 to 77.6) (OR)
Hancock et al 2017	2 findings present	0 findings	Days to recurrence of activity limiting LBP (1 y)	5.9 (0.8 to 44.8) (HR)
Kasch et al 2021b	3 MRI findings	No MRI finding	Pain (NRS) (6 y)	4.9 (–2.3 to 12.2)
Hancock et al 2017	3 findings present	0 findings	Days to recurrence of activity limiting LBP (1 y)	24.0 (1.1 to 518.6) (OR)
Hancock et al 2017	3 findings present	0 findings	Days to recurrence of activity limiting LBP (1 y)	12.2 (1.3 to 118.2) (HR)
Kasch et al 2021b	4 MRI findings	No MRI finding	Pain (NRS) (6 y)	7.7 (0.3 to 15.2)
Kasch et al 2021b	5 or more MRI findings	No MRI finding	Pain (NRS) (6 y)	12.1 (0.4 to 23.7)
		Participants with a mixed population sample		
Carragee et al 2006	Canal stenosis (moderate/severe)	No canal stenosis	Modified ODI (5 y)	2.1 (0.9 to 5.1) (OR)
Carragee et al 2006	Canal stenosis (moderate/severe)	No canal stenosis	Serious LBP (NPRS ≥ 6 for ≥ 1 wk) (5 y)	2.90 (p = 0.09) (OR)

Carragee et al 2006	Disc degeneration (Grade 5)	No disc degeneration	Serious LBP (NPRS ≥ 6 for ≥ 1 wk) (5 y)	4.40 (p = 0.08) (OR)
Maatta et al 2015	Disc degeneration (summary score 0 to 60)	No disc degeneration	Disabling LBP lasting > 1 mo (10 y)	1.1 (1.1 to 1.1) (OR)[o]
Sääksjärvi et al 2020	Disc degeneration	No disc degeneration	VAS (30 y)	–0.6 (–2.7 to 1.5)
Sääksjärvi et al 2020	Disc degeneration (MRI group)	No disc degeneration	VAS (30 y)	–0.3 (–2.3 to 1.7)
Carragee et al 2006	Disc degeneration	No disc degeneration	Modified ODI (5 y)	1.1 (0.5 to 2.3) (OR)
Sääksjärvi et al 2020	Disc degeneration	No disc degeneration	ODI (30 y)	–5.0 (–16.7 to 6.7)
Sääksjärvi et al 2020	Disc degeneration (MRI group)	No disc degeneration	ODI (30 y)	–2.9 (–14.0 to 8.2)
Carragee et al 2005	Disc herniation	No disc herniation	Remissions of LBP for > 6 mo (VAS, 0 to 10) (63 mo)	4.81 (p = 0.01) (OR)
Carragee et al 2006	Modic change (moderate/severe)	No Modic	Serious LBP (NPRS ≥ 6 for ≥ 1 wk) (5 y)	2.50 (p = 0.10) (OR)
Maatta et al 2015	Modic type 1, 2 or 3)	No Modic	Disabling LBP lasting > 1 mo (10 y)	1.6 (1.0 to 2.4) (OR)[o]
Maatta et al 2015	Modic type 1, 2 or 3)	No Modic	Disabling LBP lasting > 1 mo (10 y)	2.2 (1.4 to 3.3) (OR)
Maatta et al 2015	Schmorl Node	No Schmorl node	Disabling LBP lasting > 1 mo (10 y)	1.1 (0.7 to 1.8) (OR)[o]

Positive outcome for difference in means indicates that exposure group had a worse outcome at follow-up compared with comparison group
Odds ratio (OR) > 1 indicates greater odds of poor outcome in exposure group compared with comparison group
Hazard ratios (HR) > 1 indicates greater incidence of outcome in exposure group compared with comparison group
Relative risk (RR) > 1 indicates that the poor outcome is more likely to develop in the exposure group
GPE = global perceived effect; LBP = low back pain; NPRS = numeric pain rating scale; NQ = Nordic questionnaire; NRS = numeric rating scale; ODI = Oswestry Disability Index; PFI = Pain Frequency Index; RMDQ = Roland Morris Disability Questionnaire; SF-36 = 36 Item Short Form Survey; VAS = visual analogue score; ? = The confidence interval reported in this study was implausible.
[a] Z score

[b] Adjusted for race, sex, and pain distribution (ie, patients with LBP vs those with radiculopathy)
[c] Adjusted for age, sex, race, marital status, compensation, smoking status, herniation location, working status, stomach comorbidity, depression, diabetes, other comorbidity, self-rated health trend, duration of most recent episode, treatment preference, baseline score (for SF-36, Oswestry Disability Index, and Sciatica Bothersomeness Index), and centre
[d] Beta coefficient
[e] Adjusted for age, age squared, BMI, history of smoking, Disc degeneration summary score, Schmorl Node presence, and Modic change at bassline
[f] Adjusted for age, gender and early leg pain resolution (< 6 mo)
[g] Adjusted for age, attitude and beliefs about back pain, acute pain at baseline, neurological symptoms of legs, presence of continuous back pain
[h] Adjusted for age, gender, and early leg pain resolution (< 6 mo)
[i] Pearson's *r*
[j] Adjusted for age, gender, centre, race, marital status, smoking status, BMI, work status, health insurance status, compensation, joint problems, migraines, neurologic deficit, baseline back pain score, baseline satisfaction with symptoms, self-rated health trend, herniation (level, location, and morphology for combined herniation type analysis)
[k] Adjusted for baseline value of the outcome, age, sex, smoking, BMI
[l] Adjusted for radiculopathy, age, sex, and intervention group
[m] Adjusted for radiculopathy, age, sex, intervention group, tender points, baseline pain score, worrying and health anxiety, fear avoidance, low alcohol use, and exercise in leisure time
[n] Adjusted for treatment and interaction between imaging findings and treatment
[o] Adjusted for each other (disc herniation and nerve root compression)
[p] Adjusted for all variables in model, demographics, and care pathway
[q] Adjusted for age, gender, and baseline symptoms
[r] Adjusted for age, gender, and size of Modic change at baseline

Chapter Seven:

Exploring study level factors which potentially explain differences in results of studies evaluating the STarTBack stratified treatment approach for low back pain: A literature review

Statement from co-authors confirming authorship contribution of the PhD candidate

The co-authors of the paper *"Exploring study level factors which potentially explain differences in results of studies evaluating the STarTBack stratified treatment approach for low back pain: a literature review"* confirm that Christopher S Han has provided the following contributions to the study:

- Conception and design of the research
- Data acquisition
- Data analysis and interpretation of findings
- Writing of the manuscript and critical appraisal of the content

Christopher S Han _____ Date: 22/11/2023

Professor Chris G Maher _____ Date: 22/11/2023

Professor Mark J Hancock _____ Date: 22/11/2023

Exploring study level factors which potentially explain differences in results of studies evaluating the STarTBack stratified treatment approach for low back pain: A literature review

Han CS[1], Hancock MJ[2], Jones CM[1], Maher CG[1]
1 The Institute for Musculoskeletal Health, The University of Sydney and Sydney Local Health District, Sydney, Australia.
2 Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia.

Correspondence (for review) for contact purposes only:

Christopher S Han

Medicine and Health, School of Public Health, University of Sydney

Institute for Musculoskeletal Health

Australia

+61 403 005 696

christopher.han@sydney.edu.au

Correspondence (for publication) for contact purposes only:

Christopher S Han

Medicine and Health, School of Public Health, University of Sydney

Institute for Musculoskeletal Health

Australia

christopher.han@sydney.edu.au

Abbreviated title: Differences in effects of the STarTBack Tool: A literature review

Key words: STarTBack, low back pain, explanatory, pragmatic, implementation

Word count: 443 (Abstract), 4576 (Introduction, Methods, Results, Discussion, Conclusion)

References: 19

Tables: 2

Figures: 2

Ethics and dissemination: Ethical approval is not required.

Conflict of interest: None declared.

Source(s) of support: None declared.

Acknowledgements: None declared.

Correspondence: Christopher S Han

Funding source: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Abstract

Objective: Are there differences in the design and/or conduct of studies that have tested the STarTBack treatment approach for the management of LBP, potentially explaining differences in study results?

Design: Literature review.

Setting: Primary Care.

Participants: Participants with LBP.

Results: Eleven studies conducted across 5 countries were included. There were substantial differences in the proportion of patients allocated to the different risk groups; low-risk group (range: 19% to 58%), medium risk (range: 31% to 52%), and high-risk (range: 6% to 38%). There were large differences between studies in the implementation of the STarTBack approach. The original STarTBack trial (Hill et al 2011) had a more explanatory design while in many subsequent studies, the design was more pragmatic/real world. Only the two original studies provided clear evidence that the implementation of the STarTBack tool led to a higher proportion of patients receiving matched treatment. In the other studies there was no evidence of a difference, or it was unclear. In two studies a researcher made the decision about which matched treatment patients received based on the STartTback Tool, while in all the other studies this was done by a clinician. Most studies recommended the same matched treatment for each risk group as per the original study except for a small number of studies. Only three studies reported whether the clinician delivering matched treatment followed the recommended treatment as per the tool. There was substantial variability in the training clinicians received.

Conclusions: Reporting of important study level factors in how the STarTBack approach was implemented was unclear. There is some suggestion that key factors may include the individual who implemented the STarTBack tool, whether the recommendations of the tool were followed, the amount of training the clinician delivering the matched treatment received, and whether clinicians actually delivered the matched treatment.

Review registration: OSF Registries: https://doi.org/10.17605/OSF.IO/QJZXY

Strengths and limitations of this study

- Strengths of this study include using a sensitive search strategy and following a strict pre-specified protocol.
- We searched various databases and included several different study designs[16], which allowed us to summarise the available literature on the study level factors that may explain the differences in findings for studies using the STarTBack approach.
- A limitation of our review is that many studies did not clearly report on the study level factors we explored.
- Due to the qualitative nature of the study level factors explored, we aimed to mitigate misinterpretation using two independent reviewers and a third reviewer were necessary.

Introduction

Patients with low back pain often seek treatment in primary care from clinicians such as a general practitioner (GP) or physiotherapist.[1] There is currently limited evidence to guide clinicians to shift from generic treatment and move towards precision or personalised medicine approaches.[1] One such approach that has been investigated in a number of studies is risk-based stratified care.

Risk-based stratified care involves selecting treatments according to a patients' risk of persistent disabling pain, to maximise treatment benefit and reduce potential harm and wasted resources associated with unnecessary interventions.[2] The STarTBack approach is a widely known and investigated example of a risk-based stratified care approach, introduced by Hill and colleagues in 2008.[3] The STarTBack approach was originally developed with the intention to match patients to specific treatment approaches based on prognosis or risk of poor clinical outcomes.[4] The initial randomised controlled trial (RCT) by Hill et al 2011[5] (n=851) and the follow up pre-post design study by Foster et al 2014[6] (n=922) assessing the effectiveness of the STarTBack approach compared to non-stratified therapist led care, demonstrated initial promise. The Hill et al 2011[5] and Foster et al 2014[6] studies reported that the STarTBacK approach resulted in slightly better patient outcomes (effect on 0-100 scale: 5.5 points (95% CI: 2.3 to 8.6) at 4 months follow-up, and more efficient use of resources in primary care patients with low back pain.

Since the initial studies from Hill et al 2011[5] and Foster et al 2014[6], subsequent trials[1,7-15] investigating stratified care using the STarTBack Tool have failed to replicate the initial positive results. For example, Morsø et al 2021 performed a randomised trial and found no

difference in pain or disability outcomes between low back pain patients who received stratified care compared to non-stratified care.[13]

The failure of subsequent studies[1,7-15] to replicate the positive results of the initial STarTBack studies[5,6] may potentially be explained by differences in the design and/or conduct of subsequent studies. For example the Hill et al 2011 trial[5] was an explanatory trial testing the effectiveness of the STarTBack approach under ideal conditions and was less reflective of routine clinical care (although still an important study design). This could largely explain the high fidelity, and therefore may explain the small but significant outcome in favour of a stratified approach in this study.[5] However, in subsequent studies such as Delitto et al 2021[10], the STarTBack tool was used to generate recommended treatments, however this was used as an opportunity for clinicians and patients to discuss the most appropriate treatment option. Intervention fidelity was not recorded in this study[10], however, the way the STarTBack tool was used may be more reflective of what occurs in clinical practice and the challenges associated with implementing stratified approaches. Exploring the differences between studies will help clinicians and researchers understand how best to implement the STarTBack Tool (if at all) and may inform the development and testing of other complex models of care. Given the conflicting study results regarding the effectiveness of the STarTBack approach and no studies having explored factors potentially influencing these differences, we conducted this literature review to qualitatively explore any differences.

Objective

Primary objective

To explore differences in study level factors in studies that have tested the STarTBack treatment approach for the management of musculoskeletal pain, and to investigate whether they may explain differences in study results.

Methods

A review protocol was specified in advance and registered on OSF Registries: https://doi.org/10.17605/OSF.IO/QJZXY

Search strategy

MEDLINE, CINAHL, EMBASE were searched from inception to 26th July 2023. The search strategy included a combination of several terms for each of three domains: low back pain, stratification, and randomised controlled trials. The full search strategy is presented in Supplementary Appendix S1. We also screened reference lists of included studies for additional studies. Forward citation searching was also performed. We included studies in any language.

Inclusion criteria

Eligible studies met the following criteria:

i) randomised controlled trials (RCTs), controlled clinical trials, and interrupted time series[16] designs
ii) used the STarTBack Tool to stratify participants into sub-groups
iii) studies provided matched treatments according to participants STarTBack score (even if a different matched treatment to that originally proposed by Hill et al 2011[5] was used)

One review author (CSH) excluded clearly irrelevant titles. Two review authors (CMJ and CSH) independently screened abstracts to exclude irrelevant studies. Two reviewer authors

(CMJ and CSH) then independently reviewed full texts for eligibility based on the inclusion criteria. Any disagreements were resolved through discussion and a third reviewer (CGM or MJH).

Study level factors that were qualitatively explored

- Who administered the STarTBack Tool to the patient?
- Did the implementation of the STarTBack Tool lead to changes in the recommended or received interventions?
- Who made the decision about which matched treatment patients received based on STarTBack score?
- Which clinician provided matched treatments?
- Did the clinician delivering matched treatment follow the recommendations for treatment based on the tool?
- Was the clinician implementing the matched treatment trained to do so?
- Treatment received in the control group

Clinical outcomes

To assess the effect of each study level factor that may explain the differences in study results, we extracted pain and disability outcomes at all available time points.

Data extraction and analysis

Two review authors independently reviewed the search results and extracted data into the data extraction form. A form was used to record characteristics of the study (e.g., study design, healthcare setting, participant age, follow-up duration, duration of pain) and study level factors as listed above.

Due to the exploratory nature of this study, results are presented descriptively. Data relevant to (but not limited to) study design used, the healthcare setting the of each study (including country), and characteristics of the patients (e.g., age, sex, duration of pain) were extracted.

We defined community clinicians (e.g., GP, physiotherapist) as clinicians who practise in the community as part of their usual care. We defined a research clinician (e.g., GP, physiotherapist) as clinicians working as part of a research team solely for the purpose of the research project.

Results

Characteristics of included studies

Our search retrieved 5399 records. After removing duplicates (n=190) 5209 records were identified. Of these, 4866 were clearly irrelevant based on title screening. A further 323 records were excluded based on title and abstract screening. Following full text review by two independent authors, 10 studies were included (Figure 1). One additional study was found by contacting experts in the field, resulting in a total of 11 included studies. Excluded studies and primary reasons for exclusion are presented in Supplementary Appendix S2.

Characteristics of included studies

Eleven studies were included.[5-15] Sample sizes ranged from 109 to 2971 (median: 550), studies were conducted in 5 different countries (3 in the UK, 5 in the USA, 1 in Denmark, 1 in Netherlands, and 1 in Ireland), and were all conducted in primary care. Nine studies included participants with low back pain and one study included patients with sciatica[11]. One study

included participants with neck and low back pain.[9] Follow-up ranged from 4 weeks to 12 months. There was substantial between-study variability of the proportion of patients allocated to the different risk groups. The proportion of patients in the low-risk group ranged from 19% to 58%, in the medium risk group 31% to 52%, and in the high-risk group from 6% to 38%. One study[9] did not report the results of the proportion of patients in each risk group. Study level variables are presented in Table 1 and further details are presented in Supplementary Appendix S3.

Study population and protocol for matched treatment

Patient population data were reported in all eleven studies and are presented in Table 1.[5-15] Most studies recruited similar participants to the original studies[5,6], including participants with low back pain with or without leg symptoms. However, in one study[11], only participants with sciatica were recruited which resulted in altered matched treatment received in this study compared to the other included studies. In another study the medium and high-risk patient groups both received the same matched treatment and participants were combined and reported in total[9], therefore did not follow matched treatment as per the original STarTBack approach. One study[10] only included patients who scored as high-risk on the STarTBack Tool, however the matched treatment provided followed the STarTBack approach. All other studies[5-8,11-15] used the same STarTBack approach to matched treatment as the original studies (Hill et al 2011[5] and Foster et al 2014[6]). A flow diagram of how the STarTBack Tool is implemented and where the main differences between studies in the tool's implementation is presented in Figure 2.

Study design and healthcare setting

Four studies[5,11,13,15] were randomised controlled trials (RCT) and five studies[7-10,12] were cluster RCTs. One study[14] was a non-randomised controlled trial and one study[6] was a sequential comparison.

Who administered the STarTBack Tool to the patient

Data on who administered the *STarTBack* tool to the participant were available in 8 of the 11 studies.[5-8,11,13-15] In the original Hill et al 2011[5] study, a research clinician (i.e., a research nurse) administered the tool, and in three other studies[8,11,13] a member of the research team (i.e., research specialist, project secretary, research clinician) administered the tool. The community general practitioner (GP) administered the tool in one study (Foster et al 2014)[6] and a community physiotherapist administered the tool in one study[14]. The tool was self-administered by the patient in two studies[7,15] and it was unclear in three studies[9,10,12] who administered the tool.

Who made the decision about which matched treatment patients received based on STarTBack score?

Data on who made decisions about the matched treatment based on the use of the tool were available in 8 of the 11 studies.[5-8,10,11,13,15] In two studies[5,11] including the original Hill et al 2011[5] study, a member of the research team (i.e., research physiotherapist) made the decision regarding the patient's treatment, based on the tool. The community general practitioner (GP) made decisions about treatment using the tool in one study (Foster et al 2014)[6], and a community physiotherapist made the decision about patient treatment in one study[7]. In one study, a community clinician (i.e., physicians, nurse practitioners, or physicians' assistant) made decisions about treatment using the tool.[15] In three studies[8,10,13], patients were stratified through a variation of electronic health database record systems based on results of the tool,

however, it was unclear who made the final decision regarding the patient's treatment. It was unclear in three studies[9,12,14] who made the decision about the patient's treatment based on the tool results.

Did the implementation of the STarTBack Tool lead to changes in the clinicians recommended interventions?

Data on whether clinician behaviour (recommended treatment) changed with the use of the STarTBack tool were reported in only 3 of the 11 studies.[5,6,10] In the original Hill et al 2011 study[5], implementation of the STarTBack tool led to clinicians following the recommended matched treatment based on the risk matched recommendations from the tool in 97% of patients, and led to higher referral to physiotherapy in the intervention/stratified group (75%) compared to the control group (58%). In the Foster et al 2014 study[6], implementation of the STarTBack tool led to clinicians following the recommended matched treatment based on the risk matched recommendations from the tool in 71% of patients, and led to higher referral for physiotherapy in the medium and high-risk groups in the intervention/stratified group (72%) compared to the control phase (40%). In another study[10], the implementation of the tool led to higher referral for physiotherapy in the high-risk group in the intervention/stratified group (57%) compared to the control phase (30%). However, in this study[10] it was not reported if clinicians followed the recommended treatment based on the risk matched recommendations from the tool. In eight studies[7-9,11-15] it was unclear or not reported as to whether the recommended treatment matched the tool's recommendation, or if the implementation of the STarTBack tool led to changes in the treatment received.

Which clinician provided matched interventions?

Data on who provided the matched treatments based on their STarTBack score were available in 10 of the 11 studies.[5-7,9-15] In the original Hill et al 2011 study[5], a research physiotherapist provided matched treatment for one session (for low risk treatment) at baseline, and then medium and high risk patients were referred for further community physiotherapy.

For patients stratified to low-risk treatment, a research physiotherapist provided matched treatment in one study[11], a community general practitioner provided matched treatment in one study (Foster et al 2014)[6], a community physiotherapist provided matched treatment in five studies[7,12-15], and in one study[9] a community physiotherapist and 'spine coach' (professionally trained in motivational interviewing) provided matched treatment. It was unclear in one study[8] which clinician provided matched treatment (e.g., either a GP, nurse, physiotherapist) and in another study[10] only results for high-risk participants were presented, which are discussed below.

For patients stratified as medium risk, community physiotherapists provided matched treatment in six studies[6,7,12-15], a research physiotherapist provided matched treatment in two studies[5,11], and in one study[9] a community physiotherapist and 'spine coach' (professionally trained in motivational interviewing) provided matched treatment. It was unclear in one study[8] which clinician provided matched treatment (e.g., a GP, nurse, or physiotherapist). One study[10] only presented results for high-risk participants, where a community physiotherapist provided matched treatment.

For patients stratified as high risk, a community physiotherapist provided matched treatment in seven studies[6,7,10,12-15], a research physiotherapist provided matched treatment in one study (Hill et al 2011)[5], and in another study[9] a community physiotherapist and 'spine coach'

(professionally trained in motivational interviewing) provided matched treatment. In one study[11], including only patients with sciatica, high risk participants were referred immediately for MRI and to a Specialist, differing to the original (Hill et al 2011)[5] STarTBack approach. It was unclear in one study[8] which clinician provided matched treatment (e.g., either a GP, nurse, or physiotherapist).

Was the clinician implementing the matched treatment trained to do so?

Data on training of the clinician providing the matched treatment were available in 9 of the 11 studies.[5-8,10,12-15] In the original Hill et al 2011[5] study, clinicians providing medium risk treatment received 4 days of training and 10 days of training for high risk treatment. One study[8] trained clinicians (e.g., physicians, nurses) in delivering matched treatment over 6 sessions (1 hour each) over a 6-month period and physiotherapists received 5 days of training for matched treatments for medium to high-risk groups. One study[15] trained community physiotherapists over a 2-day course (hours not specified) to provide matched treatments for all risk groups. One study[10] trained clinicians for a total of 10.5 hours in delivering matched interventions only for high-risk groups. In one study[13] physiotherapists received 5 days of training, additional booster sessions, and access to consulting physiotherapists (trained in STarTBack approach) to provide matched treatment for all risk groups. In another study[7], physiotherapists were trained (up to 12 hours over 4 weeks) in the implementation of the matched treatments for all risk groups. Two studies[6,14] reported that physiotherapists were trained, however, did not report the amount of training provided. In two studies[9,11] it was not reported whether training was provided to clinicians on risk matched treatment and in another study[12] it was unclear if the training provided was specific for providing matched treatment.

Did the clinician delivering matched treatment follow the recommendations for treatment based on the tool?

Data on whether treating clinicians delivering matched treatment followed the recommended treatment (as per each study protocol for matched treatment) were available in 3 of the 11 studies.[5,6,11] In the original Hill et al 2011[5] study, among the participants referred for physiotherapy, initial treatment attendance was reported as 93% in the intervention group (i.e., STarTBack group). No data were reported based on each risk group in either of the two studies.[5,6] In one study[11] clinicians followed the recommended treatment in the low-risk group in almost all patients (except 1 patient), in the medium risk group recommended treatment was followed in 82.5% of patients, and in the high-risk group recommended treatment was followed in 81.3% of patients. In the three studies[5,6,11] reported above, the reasons for patients not receiving the recommended matched treatment based on the tool were not clear or not reported. In eight studies[6-8,10,12-15] it was unclear or not reported whether clinicians followed the recommended matched treatment based on the tool. In one study[9], patients in the medium and high-risk groups received the same treatment and data on whether clinicians followed the recommended treatment based on the tool was not reported.

Treatment received in the control groups

Data on the intervention received in the control group were available in 9 of the 11 studies.[5-7,9,11-15] In all eleven studies the STarTBack score was collected, however all control groups did not use the STarTBack Tool to determine treatment or the results of the tool were not available to the clinicians providing treatment. In the original Hill et al 2011 study[5], decisions about referral in the control group were made based on the physiotherapists' clinical judgment, without knowledge of a patient's STarT Back Tool classification. In four studies[5,7,12,14] control participants received some form of intervention by a community or research physiotherapist

(e.g., treatment based on guidelines). In four studies[6,11,13,15], control participants received different forms of interventions from a community GP or primary care clinician (i.e., physicians, nurse practitioners, and physicians' assistant) and they were advised to provide treatment as usual (e.g., referral for physio, medication, or guideline-based treatment). In one study[9] there were two control groups, one group received an individualised exercise program (not based on risk score), and the other group received no treatment. In two studies[8,10] it was unclear what treatment was received in the control group.

Discussion

Principal findings

This review found that reporting of implementation of the STarTBack approach was often unclear for many important study level factors. However, there were important differences between studies in the implementation of the STarTBack approach, which may explain differences in results between studies. The original STarTBack RCT (Hill et al 2011[5]) had a more explanatory design while in many of the subsequent studies, the design was more pragmatic/real world. For example, in the original STarTBack RCT (Hill et al 2011[5]) the tool was administered by a researcher whilst in two studies[6,14] the tool was administered by a community clinician (which is more reflective of clinical practice). In most studies it was unclear, or not reported, if the implementation of the STarTBack tool resulted in (between group) differences in the recommended or received interventions. Only three studies[5,6,10] (including the original Hill et al 2011 study) provided clear evidence that implementation of the STarTBack tool led to a higher proportion of patients receiving matched treatment. In the other studies there was either no evidence of a difference, or it was not clear. Only three studies[5,6,11] reported whether the clinician responsible for delivering matched treatment followed the recommended treatment. Most studies reported that training was provided to

clinician however, there was a large range in the amount of training provided (range: 6 hours to 10 days). The amount of training was reported in only half of the included studies. In all studies, the intervention received in the control group either did not include use of the STarTBack Tool or the risk profile based on the tool was not available to treating clinicians.

Strengths and weaknesses of the study

Strengths of this study include using a sensitive search strategy and following a strict pre-specified protocol. We searched various databases and included several different study designs.[16] This allowed us to summarise the available literature on the study level factors that may explain the differences in findings for studies using the STarTBack approach. We were able to identify key differences between studies that may be useful for future research and how clinicians may implement this tool in clinical practice. A limitation of our review is that many studies did not clearly report on the study level factors we explored. Due to the qualitative nature of the study level factors explored, we aimed to mitigate misinterpretation using two independent reviewers and a third reviewer were necessary. We did not assess risk of bias for any of the included studies in our review and did not approach this review with a quantitative approach as commonly performed in systematic reviews, as we anticipated heterogeneity in the availability and reporting of data.

Comparison with previous studies

To our knowledge this is the first review exploring the potential study level factors that may explain the differences in study outcomes. Two systematic reviews[17,18] have previously investigated the effectiveness of stratified care compared to non-stratified care (including the STarTBack approach). One review (n=8 studies) included four randomised controlled trials and 4 non-randomised controlled trials[17] and another review (n=24 studies)[18] included the

STarTBack approach as part of other stratification systems. Neither study[17,18] directly answers our research question and no study to date has explored potential factors that may contribute to the differing study results.

A recent Journal of Physiotherapy editorial discussed potential reasons behind the different outcomes between studies investigating the effectiveness of the STarTBack approach.[19] The editorial considered one particular explanation for the lack of consistency in study results compared to the original studies: intervention fidelity. The editorial highlighted the importance of the two key components of stratified care (i.e., subgrouping patients based on risk scores and providing targeted matched treatments).[19] All studies in our review sub-grouped participants based on their risk scores, however not all participants went on to receive matched interventions. Our review highlighted that it was often not clearly reported whether participants received matched treatments (e.g., only 3 of 11 studies providing data on whether patients received matched care), limiting our ability to explore this further. Reporting on whether participants received matched care is critical to interpreting the results of studies as it likely contributes to the different outcomes in studies investigating the implementation of the STarTBack Tool. The effectiveness of the STarTBack approach is also likely dependent on the amount of training clinicians delivering matched care received, and therefore had the clinical skills to provide the matched treatments. There was substantial variability in the amount of training provided to clinicians (range: 6 hours to 10 days) and in half the studies it wasn't clear the amount of training clinicians received or if training of clinicians was even provided.

Implications for future research

It may be worthwhile for future research investigating stratified care (i.e., STarTBack approach) to use relatively explanatory research designs like the original Hill et al 2011[5] study

to determine if there is merit in continuing with subsequent pragmatic studies. Following the initial Hill et al 2011 study[5], very few, if any, explanatory studies have been completed. Important nuances on implementing the STarTBack approach and feasibility may have been missed by performing future studies based on a lack of available research. If future studies demonstrate feasibility, are able to appropriately subgroup patients based on risk scores, and provide targeted matched care, larger and more pragmatic randomised controlled trials should be conducted to assess the clinical effectiveness of the STarTBack approach. Future studies should also focus on how best to implement the STarTBack approach and strategies to monitor and improve the implementation of the tool should also be considered. A better understanding of the key variables likely influencing the implementation of matched interventions may help inform clinicians of the best approach to implement stratified care in clinical practice. A related issue to implementing the STarTBack approach is whether participants received the matched intervention they were matched to by the STarTBack tool and whether clinicians implementing matched care had the appropriate training and skills to do so. Future research should better monitor whether participants go on to receive the intervention they are matched too, and more importantly, investigate the potential reasons for why patients did not receive matched treatment e.g., cost, access, clinician training.

Conclusions

We found that in most studies important study level factors in relation to the implementation of the STarTBack approach (use of the tool and matched care) were not clearly reported. As a result, it was difficult to determine which factors may have impacted on study outcomes. There is some suggestion that the key factors might be the individual who implemented the STarTBack tool, whether the recommendations of the tool were followed, the amount of training the clinician delivering the matched treatment received, and whether they actually

delivered the matched treatment. Future research on the STarTBack approach must clearly report on the key implementation element of the approach and be designed to investigate the most important elements of implementation to produce better patient outcomes.

Table 1. Characteristics of included studies

Study	Study design	Study sample/healthcare setting	Participant age	Follow-up duration	Duration of pain
Beneciuk 2015	Cluster RCT	NSLBP +/- leg pain Primary Care	IG: 46.0 (SD 12.8) CG: 46.6 (SD 11.2) Female: 58.7%	4 wks	≤2 wk to ≥ 3 mo Highest proportion of patients IG: >3 mo; 57% CG: >3 mo; 59%
Cherkin 2018	Cluster RCT	NSLBP +/- leg pain Primary Care	IG: 58.0 (SD 18.4) CG: 55.0 (SD 17.3) Female: 56%	2, 6 mo	<3 mo up to 12 mo Highest proportion of patients IG: <3 mo; 56% CG: <3 mo; 56%
Choudhry 2022	Cluster RCT	Neck and/or NSLBP Primary Care	IG: 50.9 (SD 16.0) CG1: 52.6 (SD 16.0) CG2: 51.2 (SD 16.0) Female: 60%	3, 12 mo	<3 mo Highest proportion of patients IG: <3 mo; 100% CG: <3 mo; 100%
Delitto 2021	Cluster RCT	NSLBP +/- leg pain Primary Care	IG: 49.3 (SD 16.2) CG: 50.6 (SD 17.1) Female: 59%	6, 12 mo	<1 mo to >3 mo Highest proportion of patients IG: <1 mo; 62% CG: <1 mo; 59%
Foster 2014	Sequential comparison	NSLBP +/- leg pain Primary Care	Pre: 53.0 (SD 15.0) Female: 55% Post: 54.1 (SD 14.8) Female: 60%	6 mo	<1 mo to >3 yrs Highest proportion of patients Pre: 6 mo to 3yrs; 25% Post: 6 mo to 3yrs; 24%
Hill 2011	RCT	NSLBP +/- leg pain Primary Care	IG: 50.1 (SD 15) CG: 49.1 (SD 14.3). Female: 500/851	4, 12 mo	<1 mo to >3 yrs Highest proportion of patients IG: 6 mo to 3yrs; 25% CG: 1 to 3 mo; 23%
Konstantinou 2020	RCT	Sciatica Primary Care	IG: 50.7 (SD 14.5) CG: 53.3 (SD 13.5) Female: 55%	4, 12 mo	<2 wk to >12 mo Highest proportion of patients IG: 2-6wk; 42% CG: 2-6wk; 41%
Koppenaal 2022	Cluster RCT	NSLBP, unclear if leg pain present Primary Care	IG: 48.1 (SD 15.1) CG: 47.3 (SD 13.6) Female: 49%	3 mo	0 wks to >12 mo Highest proportion of patients IG: >12 mo; 45% CG: 0 to 6wk; 47%
Morsø 2021	RCT	NSLBP, unclear if leg pain present Primary Care	IG: 46 (35-57)* CG: 45 (31-57)* Female: 58%	3, 12 mo	<1 mo to > 12 mo Highest proportion of patients IG: >12 mo; 36% CG: <1 mo; 32%
Murphy 2016	Non-randomised controlled trial	NSLBP +/- leg pain Primary Care	IG: 43.4 (SD 11.4) CG: 43.1 (SD 10.7) Female: 71%	3 mo	> 3 mo Highest proportion of patients IG: >3 mo; 100% CG: >3 mo; 100%
Rhon 2023	RCT	NSLBP +/- radiculopathy Primary care	IG: 34.7 (SD 8.1) CG: 33.6 (8.8) Female: 34%	6 wk, 6 mo, 12 mo	Any pain duration in both groups

IG=Intervention group; CG=Control group; NSLBP=Non-specific low back pain; RCT=Randomised controlled trial; SD=Standard deviation
*range

Table 2. Study level implementation variables

Study	Individual administering tool	Trained to use tool	Clinician providing matched treatment?	Did the tool lead to changes in the recommended treatment?	Proportion receiving matched treatment?	Allocated treatment
Beneciuk 2015	Clinician (Physiotherapist)	Yes	All risk groups: Community physiotherapist	Not reported	Unclear	IG: Treatment based on STarTBack score CG: Physiotherapy as 'usual'
Cherkin 2018	Researcher	Yes	All risk groups: A PCP	Not reported	Unclear	IG: Treatment based on STarTBack score CG: "Usual care"
Choudhry 2022	Unclear	Not reported	All risk groups: Physiotherapist, "spine coach" and an "ICE MD"	Not reported	Not reported	IG: Treatment based on STarTBack score CG1: IPT CG2: No intervention
Delitto 2021	Unclear	Yes	High risk: Community physiotherapist	Higher referral for Physiotherapy in IG (high risk group) (57% vs 30%)	Not reported	IG: Treatment based on STarTBack score CG: 'Usual care'
Foster 2014	Clinician (GP)	Yes	Low risk: Community GP Medium and high risk: Referral to community physiotherapist	Higher referral for physiotherapy in medium and high-risk group (72% vs 40%) No difference for low-risk group (65% vs 68% not referred)	Clinician followed recommended matched treatment in 71% of patients[a]	IG: Treatment based on STarTBack score CG: GP +/- physiotherapy referral if appropriate
Hill 2011	Researcher	Yes	Low risk: Research physiotherapist Medium and high risk: Research physiotherapist	Higher referral for Physiotherapy in IG (75% vs 58%)	Clinician followed recommended matched treatment in 93% of patients[a]	IG: Treatment based on STarTBack score CG: Physiotherapy referral if appropriate
Konstantinou 2020	Researcher	Not reported	Low and medium risk: Research physiotherapist High risk: Referred for MRI and Specialist	Not reported	Clinician followed recommended matched treatment in: Low risk: 98.8% Medium risk: 82.5% High risk: 81.3%	IG: Treatment based on STarTBack score *(high risk referred for MRI and to specialist)* CG: GP and Physiotherapy consult
Koppenaal 2022	Unclear	Unclear	All risk groups: Community physiotherapist	Not reported	Not reported	IG: Treatment based on STarTBack score CG: Guideline based care as per RDAP by Physiotherapist
Morsø 2021	Researcher	Yes	All risk groups: Community physiotherapist	Not reported	Not reported	IG: Treatment based on STarTBack score CG: Guideline based care by GP and Physiotherapist
Murphy 2016	Clinician (Physiotherapist)	Yes	All risk groups: Community physiotherapist	Not reported	Unclear	IG: Treatment based on STarTBack score CG: Generic physiotherapy
Rhon 2023	Self-administered	Yes	All risk groups: Community physiotherapist	Not reported	Not reported	IG: Treatment based on STarTBack score CG: "Usual care"

[a]data on whether matched treatment was followed per risk group not provided in study.
Use of the tool=who administered the tool/who stratified patients; EMR=electronic medical records; EHR=electronic health record; GP=general practitioner; PCP=primary care practitioner (e.g., either a GP, nurse, or physiotherapist); PT=physiotherapist; PIPT=psychologically informed physiotherapy; ICE MD= identify, coordinate, enhance (ICE) model of care; MD=medical doctor; IG=intervention group; CG=control group, RDAP=Royal Dutch Association for Physiotherapy; clinical physiotherapist=a primary care physiotherapist not part of the research team

Author contribution statement

Christopher S Han: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing

Mark J Hancock: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Validation; Visualization; Writing – review & editing

Caitlin M Jones: Data curation; Investigation; Methodology; Validation; Visualization; Writing – review & editing

Christopher G Maher: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Validation; Visualization; Writing – review & editing

References

1. Hill JC GS, Bromley K, Saunders B, Kigozi J, Cooper V, Lewis M, Protheroe J, Wathall S, Chudyk A, Dunn KM, Birkinshaw H, Jowett S, Hay EM, van der Windt D, Mallen C, Foster NE. Risk-based stratified primary care for common musculoskeletal pain presentations (STarT MSK): a cluster-randomised, controlled trial. *Lancet Rheumatol* 2022; **4**(9): E591-E602.
2. Hingorani AD, Windt DA, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013; **346**: e5793.
3. Hill JC, Dunn KM, Lewis M, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. *Arthritis Rheum* 2008; **59**(5): 632-41.
4. Keele Uo. START BACK: Evidence Based Implementation of Stratified Care. 2021. https://startback.hfac.keele.ac.uk/2022).
5. Hill JC, Whitehurst DG, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet* 2011; **378**(9802): 1560-71.
6. Foster NE, Mullis R, Hill JC, et al. Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. *Ann Fam Med* 2014; **12**(2): 102-11.
7. Beneciuk JM, George SZ. Pragmatic Implementation of a Stratified Primary Care Model for Low Back Pain Management in Outpatient Physical Therapy Settings: Two-Phase, Sequential Preliminary Study. *Phys Ther* 2015; **95**(8): 1120-34.
8. Cherkin D, Balderson B, Wellman R, et al. Effect of Low Back Pain Risk-Stratification Strategy on Patient Outcomes and Care Processes: the MATCH Randomized Trial in Primary Care. *J Gen Intern Med* 2018; **33**(8): 1324-36.
9. Choudhry NK, Fifer S, Fontanet CP, et al. Effect of a Biopsychosocial Intervention or Postural Therapy on Disability and Health Care Spending Among Patients With Acute and Subacute Spine Pain: The SPINE CARE Randomized Clinical Trial. *JAMA* 2022; **328**(23): 2334-44.
10. Delitto A, Patterson CG, Stevans JM, et al. Stratified care to prevent chronic low back pain in high-risk patients: The TARGET trial. A multi-site pragmatic cluster randomized trial. *EClinicalMedicine* 2021; **34**: 100795.
11. Konstantinou K, Lewis M, Dunn KM, et al. Stratified care versus usual care for management of patients presenting with sciatica in primary care (SCOPiC): a randomised controlled trial. *Lancet Rheumatol* 2020; **2**(7): e401-e11.
12. Koppenaal T, Pisters MF, Kloek CJ, Arensman RM, Ostelo RW, Veenhof C. The 3-Month Effectiveness of a Stratified Blended Physiotherapy Intervention in Patients With Nonspecific Low Back Pain: Cluster Randomized Controlled Trial. *J Med Internet Res* 2022; **24**(2): e31675.
13. Morso L, Olsen Rose K, Schiottz-Christensen B, Sowden G, Sondergaard J, Christiansen DH. Effectiveness of stratified treatment for back pain in Danish primary care: A randomized controlled trial. *Eur J Pain* 2021; **25**(9): 2020-38.
14. Murphy SE, Blake C, Power CK, Fullen BM. Comparison of a Stratified Group Intervention (STarT Back) With Usual Group Care in Patients With Low Back Pain: A Nonrandomized Controlled Trial. *Spine (Phila Pa 1976)* 2016; **41**(8): 645-52.
15. Rhon DI, Greenlee TA, Poehlein E, et al. Effect of Risk-Stratified Care on Disability Among Adults With Low Back Pain Treated in the Military Health System: A Randomized Clinical Trial. *JAMA Netw Open* 2023; **6**(7): e2321929.

16. (EPOC). CEPaOoC. What study designs can be considered for inclusion in an EPOC review and what should they be called? 2017. https://epoc.cochrane.org/resources/epoc-resources-review-authors.
17. Ogbeivor C, Elsabbagh L. Management approach combining prognostic screening and targeted treatment for patients with low back pain compared with standard physiotherapy: A systematic review & meta-analysis. *Musculoskeletal Care* 2021; **19**(4): 436-56.
18. Tagliaferri SD, Mitchell UH, Saueressig T, Owen PJ, Miller CT, Belavy DL. Classification Approaches for Treating Low Back Pain Have Small Effects That Are Not Clinically Meaningful: A Systematic Review With Meta-analysis. *J Orthop Sports Phys Ther* 2022; **52**(2): 67-84.
19. Foster NE, Hill JC, Knoop J. Are we giving stratified care a fair trial? *J Physiother* 2023; **69**(2): 65-7.

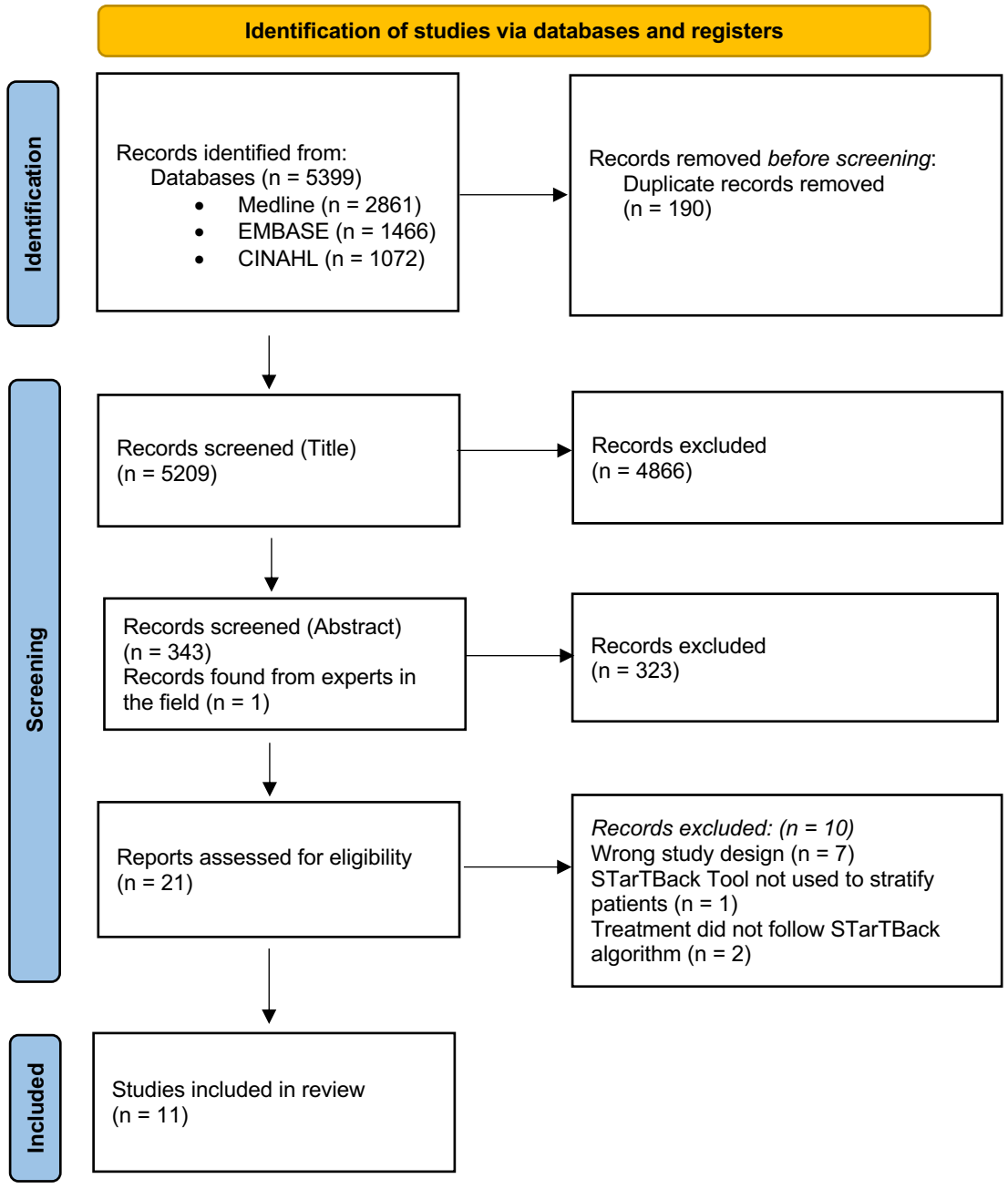


Figure 1. Study selection

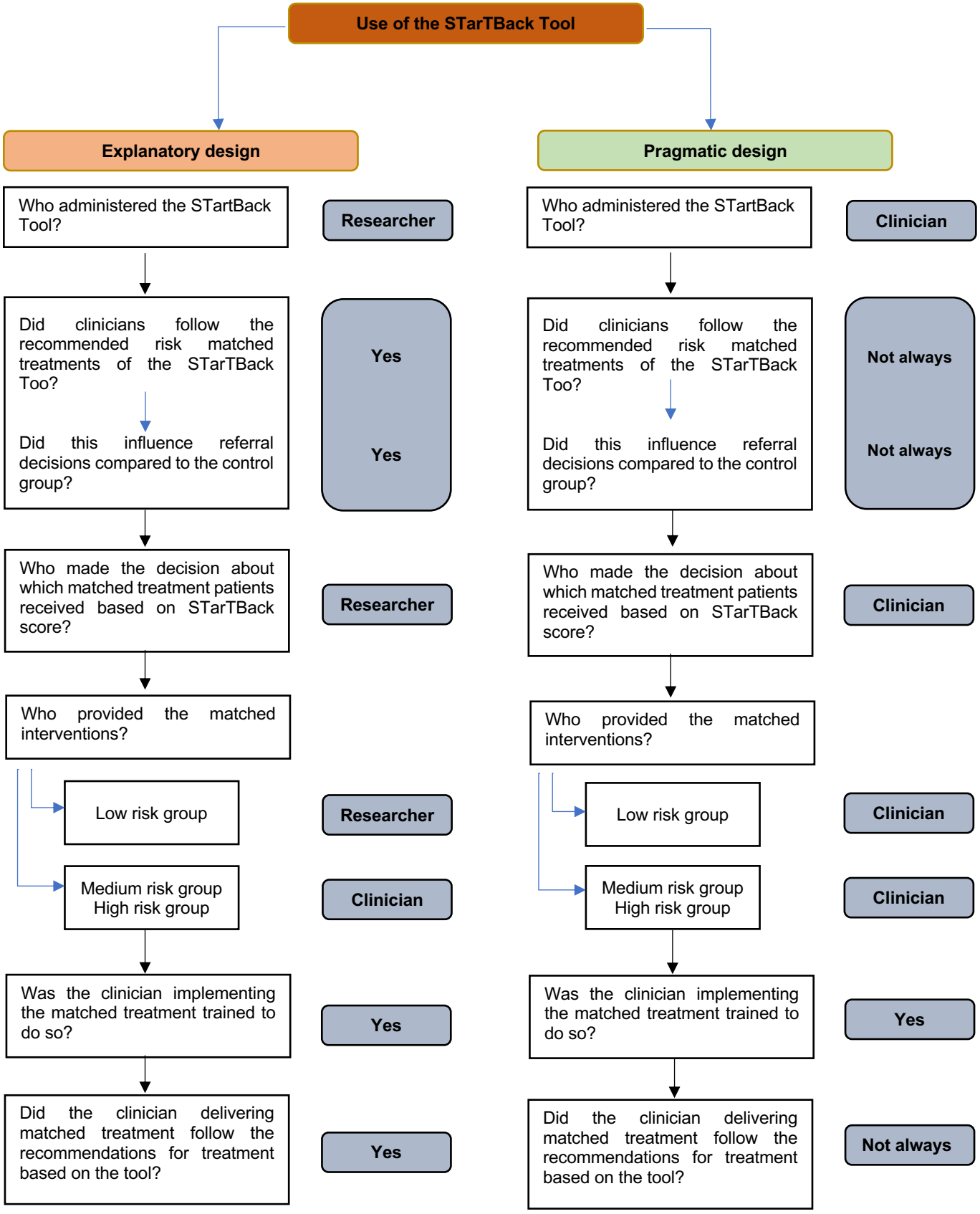


Figure 2. Flow diagram of how the STarTBack Tool is implemented in exploratory verse pragmatic study designs. Modelled off Hill et al 2011 (explanatory design) and Foster et al 2014 (pragmatic design)

Supplementary Appendix S1. Search strategy

MEDLINE		EMBASE		CINAHL	
1.	RCT.mp.	1.	RCT.mp.	1.	(MH "Back Pain+")
2.	randomised.mp.	2.	randomised.mp.	2.	MH low back pain/
3.	Randomized Controlled Trial/	3.	Randomized Controlled Trial/	3.	mh radiculopathy/
4.	randomised trial.mp.	4.	randomised trial.mp.	4.	mh sciatica
5.	cluster.mp.	5.	cluster.mp.	5.	mh "musculoskeletal pain"
6.	Prospective Studies/	6.	Prospective Studies/	6.	mh back pain or low back pain or radiculopathy or sciatica or back ache or lumbago
7.	1 or 2 or 3 or 4 or 5 or 6	7.	1 or 2 or 3 or 4 or 5 or 6	7.	S1 OR S2 OR S3 OR S4 OR S5 OR S6
8.	Back Pain/	8.	Back Pain/	8.	mh screen*
9.	low back pain/	9.	low back pain/	9.	"risk screen"
10.	radiculopathy/	10.	radiculopathy/	10.	"tool"
11.	sciatica/	11.	sciatica/	11.	"instrument/"
12.	(back pain or low back pain or radiculopathy or sciatica or back ache or lumbago).mp. [mp=title, book title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	12.	(back pain or low back pain or radiculopathy or sciatica or back ache or lumbago).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword heading word, floating subheading word, candidate term word]	12.	(mh questionnaires+)
13.	Musculoskeletal Pain/	13.	Musculoskeletal Pain/	13.	"start back/"
14.	8 or 9 or 10 or 11 or 12 or 13	14.	8 or 9 or 10 or 11 or 12 or 13	14.	"stratified care"
15.	screen$.mp.	15.	screen$.mp.	15.	TX stratif
16.	risk screen$.mp.	16.	risk screen$.mp.	16.	mh prognosis
17.	tool.mp.	17.	tool.mp.	17.	S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15 OR S16
18.	instrument.mp.	18.	instrument.mp.	18.	TX RCT
19.	questionnaire/	19.	questionnaire/	19.	"randomised/"
20.	start back.mp.	20.	start back.mp.	20.	(MH "Randomized Controlled Trials+")
21.	stratified care.mp.	21.	stratified care.mp.	21.	TX randomised trial
22.	stratif$.mp.	22.	stratif$.mp.	22.	"cluster"
23.	prognosis/	23.	prognosis/	23.	(MH "Prospective Studies+")
24.	15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23	24.	15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23	24.	S18 OR S19 OR S20 OR S21 OR S22 OR S23
25.	7 and 14 and 24	25.	7 and 14 and 24	25.	S7 AND S17 AND S24
26.	limit 25 to humans	26.	limit 25 to (human and "remove medline records")		

Supplementary Appendix S2. List of excluded full-text articles and the primary reason for exclusion

Study	Title	Reason for exclusion
Beneciuk et al 2017	Identifying Treatment Effect Modifiers in the STarT Back Trial: A Secondary Analysis	Wrong study design
Bier et al 2017	Can Primary Care for Back and/or Neck Pain in the Netherlands Benefit from Stratification for Risk Groups According to the STarT Back Tool Classification?	Wrong study design
Fritz et al 2011	Relationship Between Categorization with the STarT Back Screening Tool and Prognosis for People Receiving Physical Therapy for Low Back Pain	Treatment did not follow STarTBack algorithm
Hill et al 2022	Risk-based stratified primary care for common musculoskeletal pain presentations (STarT MSK): a cluster-randomised, controlled trial	STarTBack tool not used to stratify patients
Magel et al 2017	Outcomes of Patients with Acute Low Back Pain Stratified by the STarT Back Screening Tool: Secondary Analysis of a Randomized Trial	Treatment did not follow STarTBack algorithm
Mansell et al 2016	Exploring What Factors Mediate Treatment Effect: Example of the STarT Back Study High-Risk Intervention	Wrong study design
Middleton et al 2020	Implementing stratified care for acute low back pain in primary care using the STarT Back instrument: a process evaluation within the context of a large pragmatic cluster randomized trial	Wrong study design
Saunders et al 2020	Stratified primary care versus non-stratified care for musculoskeletal pain: qualitative findings from the STarT MSK feasibility and pilot cluster randomized controlled trial	Wrong study design
Treanor et al 2022	Prospective observational study investigating the predictive validity of the STarT Back tool and the clinical effectiveness of stratified care in an emergency department setting	Wrong study design
Whitehurst et al 2015	Implementing Stratified Primary Care Management for Low Back Pain: Cost-Utility Analysis Alongside a Prospective, Population-Based, Sequential Comparison Study	Wrong study design

Supplementary Appendix S3. All study level variables explored

Study	Individual administering tool	Clinician providing care	Did the tool lead to changes in the recommended treatment?	Proportion receiving matched treatment?	Risk level group	Allocated treatment	Was the tool collected in control group	Trained to use tool
Beneciuk 2015	Clinician (Physiotherapist)	All risk groups: Community physiotherapist	Not reported	Unclear	IG: Low (23%), Medium (44%), High (33%) CG: Low (36%), Medium (41%), High (23%)	IG: Treatment based on STarTBack score CG: Physiotherapy as 'usual'	Yes	Yes
Cherkin 2018	Researcher	All risk groups: A PCP	Not reported	Unclear	IG: Low (40%), Medium (38%), High (22%) CG: Low (42%), Medium (37%), High (22%)	IG: Treatment based on STarTBack score CG: "Usual care"	Yes	Yes
Choudhry 2022	Unclear	All risk groups: Physiotherapist, "spine coach" and an "ICE MD"	Not reported	Not reported	Unclear	IG: Treatment based on STarTBack score CG1: IPT CG2: No intervention	Yes, but has combined medium and high risk	Not reported
Delitto 2021	Unclear	High risk: Community physiotherapist	Higher referral for Physiotherapy in IG (high risk group) (57% vs 30%)	Not reported	IG: Low (36%), Medium (39%), High (24%) CG: Low (36%), Medium (40%), High (24%)	IG: Treatment based on STarTBack score CG: 'Usual care'	Yes, but only high-risk group	Yes
Foster 2014	Clinician (GP)	Low risk: Community GP Medium and high risk: Referral to community physiotherapist	Higher referral for physiotherapy in medium and high-risk group (72% vs 40%) No difference for low-risk group (65% vs 68% not referred)	Clinician followed recommended matched treatment in 71% of patients[a]	IG: Low (37%), Medium (41%), High (22%) CG: Low (39%), Medium (42%), High (20%)	IG: Treatment based on STarTBack score CG: GP +/- physiotherapy referral if appropriate	Yes	Yes
Hill 2011	Researcher	Low risk: Research physiotherapist Medium and high risk: Research physiotherapist	Higher referral for Physiotherapy in IG (75% vs 58%)	Clinician followed recommended matched treatment in 93% of patients[a]	IG: Low (26%), Medium (46%), High (28%) CG: Low (26%), Medium (46%), High (28%)	IG: Treatment based on STarTBack score CG: Physiotherapy referral if appropriate	Yes	Yes
Konstantinou 2020	Researcher	Low and medium risk: Research physiotherapist High risk: Referred for MRI and Specialist	Not reported	Clinician followed recommended matched treatment in: Low risk: 98.8% Medium risk: 82.5% High risk: 81.3%	IG: Low (22%), Medium (44%), High (34%) CG: Low (23%), Medium (45%), High (33%)	IG: Treatment based on STarTBack score *(high risk referred for MRI and to specialist)* CG: GP and Physiotherapy consult	Yes	Not reported
Koppenaal 2022	Unclear	All risk groups: Community physiotherapist	Not reported	Not reported	IG: Low (55%), Medium (36%), High (9%) CG: Low (61%), Medium (36%), High (3%)	IG: Treatment based on STarTBack score CG: Guideline based care as per RDAP by Physiotherapist	Yes	Unclear
Morsø 2021	Researcher	All risk groups: Community physiotherapist	Not reported	Not reported	IG: Low (31%), Medium (31%), High (38%) CG: Low (31%), Medium (31%), High (38%)	IG: Treatment based on STarTBack score CG: Guideline based care by GP and Physiotherapist	Yes	Yes
Murphy 2016	Clinician (Physiotherapist)	All risk groups: Community physiotherapist	Not reported	Unclear	IG: Low (17%), Medium (54%), High (29%)	IG: Treatment based on STarTBack score CG: Generic physiotherapy	Yes	Yes

Rhon 2023	Self-administered	All risk groups: Community physiotherapist	Not reported	Not reported	CG: Low (21%), Medium (50%), High (29%) IG: Low (57%), Medium (36%), High (7%) CG: Low (53%), Medium (39%), High (8%)	IG: Treatment based on STarTBack score CG: “Usual care”	Yes	Yes

Use of the tool=who administered the tool/who stratified patients; EMR=electronic medical records; EHR=electronic health record; GP=general practitioner; PCP=primary care practitioner; PT=physiotherapist; PIPT=psychologically informed physiotherapy; ICE MD= identify, coordinate, enhance (ICE) model of care; MD=medical doctor; IG=intervention group; CG=control group

Chapter Eight:
Discussion and Conclusion

8.1. Overview

The overall aim of this thesis was to investigate the role of diagnosis and targeted treatment in the management of low back pain. Specifically, this thesis explores the key issues in this field of research through (i) considering the potential misconceptions with regard to the term non-specific low back pain; (ii) investigating the diagnostic accuracy of index tests (i.e., imaging tests) available to screen for vertebral fracture; (iii) investigating the diagnostic accuracy of index tests (i.e., clinical examination and imaging tests) available to identify the pathoanatomical source of low back pain; (iv) investigating the prognostic ability of magnetic resonance imaging to predict future low back pain; and (v) investigating targeted treatments for low back pain using a classification system (i.e., STarTBack approach) and exploring differences in study level factors that may explain differences in study results.

8.2 Principal findings

The research presented in this thesis addressed key knowledge gaps with regard to diagnosis of low back pain and targeted treatments for low back pain. The study presented in Chapter Two highlighted the misconceptions with regard to the term non-specific low back and how this term has shaped the low back pain field. The term non-specific low back pain was initially intended to assist with triaging but not to form a diagnosis.[1] The term non-specific low back pain was first coined in 1956 to acknowledge that there were cases of low back pain where a specific diagnosis was not possible even following clinical examination, imaging, or laboratory tests.[2] The original proponents used the term non-specific low back pain to highlight the current challenges of diagnosis in low back pain.[1] However, they did not suggest it would remain impossible for all time to identify causes of LBP or individualise treatment[3], nor did they suggest that research into these areas should not be undertaken. For instance, Waddell's seminal paper suggested that future research should develop a rational basis for selecting the

most effective treatment for individual patients.[3] It is quite possible that misunderstandings with regard to the term non-specific low back have slowed diagnostic research and research into specific treatments for low back pain, as very little primary research of this type has been performed. There is a need to move beyond the label of non-specific low back pain through research into the diagnosis, prognosis, and specific treatment for low back pain.

The results of the systematic review and Cochrane review in Chapters Four and Five demonstrate the importance of diagnostic research in guiding how best to screen for serious low back pain and identify causes of non-serious low back pain. For those suspected of having serious low back pain, the results in Chapter Four demonstrated that there are some red flags (index tests) that can be used to guide clinicians when screening patients presenting with low back pain. However, it is important to note that the results in Chapter Four were often based on single studies with varying degrees of diagnostic accuracy. In primary care, 'trauma' demonstrated informative +LRs for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture'. 'Older age' demonstrated informative +LRs for studies in primary care for 'unspecified vertebral fracture'. 'Corticosteroid use' may be an informative red flag in primary care for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture'. A combination of index tests such as 'older age and female gender' in primary care demonstrated informative +LRs for 'unspecified vertebral fracture'. In secondary care, 'trauma' and 'older age' demonstrated informative +LRs for 'unspecified vertebral fracture' and 'osteoporotic vertebral fracture', respectively. A combination of index tests such as 'older age and trauma' in secondary care demonstrated informative +LRs for 'unspecified vertebral fracture'. When '4 of 5 tests' are positive in secondary care, they demonstrated informative +LRs for 'osteoporotic vertebral fracture'. In tertiary care, the 'presence of contusion/abrasion' was informative for 'vertebral compression fracture'. This review sheds some light on the red flags that can be used in clinical

practice; however, the challenge to identify the best red flags remains as many red flags were found too not be useful. Further research is needed to increase confidence in the diagnostic accuracy of red flags.

Forming a diagnosis for non-serious low back pain has been challenging and limited by a lack of diagnostic research. However, Chapter Five demonstrated that contrary to the belief that a diagnosis in low back pain is not possible, there are now informative diagnostic tests that increase the likelihood of the disc, sacroiliac joint, and facet joint as the source of low back pain.[4] There is now evidence that some index tests have informative likelihood ratios (+LRs and -LRs) for the disc and sacroiliac joint, and the facet joint.[4] In studies of people with persistent pain, for the disc, MRI findings of disc degeneration, HIZ, annular fissure, Modic type 1, Modic type 2, and physical examination finding of positive centralisation phenomenon increased the likelihood of the disc being the source of low back pain. For the sacroiliac joint, positive bone scan, absence of midline low back pain, and a combination of sacroiliac pain provocation tests increased the likelihood of the sacroiliac joint being the source of low back pain. For the facet joint, uptake of facet joint on SPECT increased the likelihood of the facet joint being a nociceptive source of low back pain. The evidence suggests a diagnosis may be possible for some patients with low back pain, potentially guiding targeted and specific treatment approaches or a better understanding of prognosis.

The ability to identify potential nociceptive causes of low back pain as seen in Chapter Five, is only useful if it provides a better understanding of prognosis or the natural course of low back pain. Chapter Six demonstrated that the current available evidence is insufficient to judge if most of the MRI findings we tested are, or are not, associated with future pain and disability.[5] However, the exceptions were Modic changes and disc degeneration, which were associated

with slightly worse pain and disability outcomes in the future. In populations with current low back pain the presence of Modic type 1 and 2 changes was associated with slightly worse pain and disability in the short term, the presence of Modic type 1 changes alone, was associated with slightly worse pain in the short term, and the presence of disc degeneration was associated with slightly worse pain and worse disability outcomes in the long term. In populations without low back pain, pooled estimates demonstrated the presence of disc degeneration may increase the likelihood of experiencing pain in the long term. There is some weak evidence that some nociceptive causes of low back pain identified on MRI may provide some prognostic information; however, it remains unknown if this information is useful in guiding treatment. Given the lack of high-quality evidence, it is not recommended that imaging in routine care is performed.

Another method to explore the prognosis of a patient's low back pain and provide more targeted treatments is through the utility of prognostic tools or stratification approaches such as the STarTBack approach. However, the utility and effectiveness of stratified care has been challenged in recent years.[6-14] In studies that have assessed stratified care, not all studies implemented the approach (i.e., STarTBack approach) uniformly[6-14], and this may explain why targeted matched treatments based on the tool's recommendation does not necessarily lead to better outcomes as seen in the original Hill et al 2011[15] and Foster et al 2014[16] studies. Chapter Seven demonstrated that reporting of the implementation of the STarTBack approach (use of the tool and matched treatment) was often unclear for many important study level factors, making interpretation of the results across different studies difficult. However, there were important differences in some study level factors in relation to the implementation of the STarTBack approach. The original STarTBack trial (Hill et al 2011[15]) had a more explanatory design while in many of the subsequent studies, the design was more pragmatic/real world.

Only two studies (original studies from Hill et al 2011[15] and Foster et al 2014[16]) provided clear evidence that the implementation of the STarTBack tool led to a higher proportion of patients receiving matched treatment. In the other studies there was either no evidence of a difference, or it was not clear. In two studies[9,15] (including the original Hill et al 2011 study[15]), a researcher made the decision about which matched treatment patients received based on the STartTback Tool, while in all the other studies this was done by a clinician. Most studies recommended the same matched treatment for each risk group as per the original study, but a small number of studies deviated from the original matched treatment. However, only three studies[9,15,16] reported whether the clinician delivering matched treatment followed the recommended treatment as per the tool. There was substantial variability in the training clinicians received, and in half of the studies the amount of training was not reported or was unclear.

The STartTBack Tool may improve patient outcomes when implemented under ideal conditions, however if the tool fails to change clinician behaviour or which matched treatment patients receive, or if clinicians do not deliver the correct matched treatments, then it is unlikely the tool will be effective. If this is the case, then strategies to better implement the tool (i.e., improving clinician training) need to be developed and implemented, and the tool must be re-evaluated for effectiveness under these improved conditions. The STarTBack approach needs to be explored further, as the implementation of the STarTBack approach across the literature is a limiting factor when determining the utility of stratified care in clinical practice.

8.3. Implications for practice and research

The results presented in this thesis have important implications for clinical practice and future research investigating the role of a diagnoses in low back pain. Diagnostic research is vital in order to better understand which red flags are best to screen for serious causes of low back

pain, move beyond the label of non-specific low back pain and identify potential specific causes of low back pain, and better understand prognosis and response to targeted treatments. More importantly, there is a need for more diagnostic research in the low back pain field. Further diagnostic research will help guide and inform clinicians in the future to be more selective in which patients should receive further testing to reduce unnecessary imaging, overdiagnosis, and ultimately overtreatment.

8.3.1. Implications for practice for serious low back pain

The use of red flags to screen for vertebral fractures instead of imaging all cases of low back pain still remains a clinical challenge. The available evidence highlighted in Chapter Four suggests that only four red flags are potentially useful in guiding clinical decisions regarding those who need further investigations for suspected vertebral fracture.[17]. Guidelines currently recommend the following red flags to screen for vertebral fracture (e.g., osteoporosis, history of trauma, corticosteroid use, older age, and female gender).[18] However, current guidelines do not specify which red flags may be most useful for the different types of vertebral fracture.[18] The results in Chapter Four demonstrated that trauma, older age, corticosteroid use, and presence of contusion/abrasion of informative +LRs, however confidence intervals for corticosteroid use include substantial variability. Therefore, based on the results of Chapter Four and current clinical guidelines, clinicians need to utilise a combination of red flags during the assessment, rather than using each red flag as standalone tests. Current guidelines do not report on any specific red flags for vertebral stress fracture[18] and only two studies[19,20] in our review presented data on this type of fracture. In other areas of musculoskeletal care, a combination of clinical decision rules (e.g., Ottawa ankle rules) have proven successful to reserve imaging only when suspicion of fracture is present. Given only a small number of red flags were informative, clinicians will need to utilise a combination of index tests and their

clinical expertise when screening for vertebral fracture and deciding if sufficient risk exists to order imaging. Clinicians need to consider the limitations of the research when using red flags to screen patients suspected of vertebral fracture in practice.

8.3.2. Implications for practice for non-specific low back pain

The results of Chapter Five challenge the dominant view in the low back pain field, that a pathoanatomical diagnosis is usually not possible and so the label non-specific low back pain should be used for most patients. There is now preliminary evidence that a diagnosis (e.g., origin of low back pain from disc, sacroiliac joint, or facet joint) may be possible for a subgroup of patients with low back pain.[4] However, it is important to understand that patients with low back pain may not benefit from a pathoanatomical diagnosis and a diagnosis for clinical application is not yet ready. Although the use of a diagnosis within non-specific low back pain is not ready for clinical use, it may be one element in assisting clinicians when creating a clinical profile for patients within a biopsychosocial context. For example, in Chapter Five 3 or more sacroiliac joint provocation tests were found to be informative in identifying the sacroiliac joint as the likely source of low back pain. A previous systematic review (n=5 studies)[21] also found that a cluster of pain provocation tests for the sacroiliac joint increased the likelihood of the sacroiliac joint being the source of low back pain. Clinicians may use these index tests (i.e., provocation tests for the sacroiliac joint) with confidence in clinical practice to help form the patient's clinical profile. However, clinicians should be very wary of using these diagnostic labels as there is some suggestion it may be harmful if used incorrectly.[22] It is important that treatments are not based on structure alone, but aligning clinical presentations with potential nociceptive sources could be useful for ongoing reassessment of clinical presentations to drive care. It is also important to consider issues such as feasibility, access, and costs of the index tests, and of the resulting treatment approaches they would lead to.

8.3.3. Research implications for non-specific low back pain

As there are now informative tests for the disc, sacroiliac joint, and facet joint (only one test)[4], it should be a research priority to investigate whether patients judged to have these conditions based upon validated index tests have different prognoses or responses to treatment compared to patients considered not to have those conditions. Ultimately, the diagnoses based upon these index tests will only have utility if they predict prognosis or response to treatment. Therefore, the findings from Chapter Five and Six have implications for future research.

The importance of understanding the value of diagnostic tests needs to be further evaluated in research as this will deepen our understanding on how to be more selective in which patients with non-specific low back pain are most likely to benefit from further testing. A diagnostic label will also lead to research to better understand the utility of these diagnostic tests as premature or uncritical application may lead to unnecessary imaging, overdiagnosis, and ultimately, overtreatment.[23] An important implication for research of a diagnosis is that it may also lead to the development of new treatment approaches to manage low back pain. Concurrently, a diagnostic label may also assist to select from existing treatments by understanding the mechanisms causing pathoanatomical changes or the mechanisms behind particular modes of exercises. Although a diagnosis within non-specific is not ready for clinical application, there are significant and important research implications and societal considerations.

Some of the informative index tests included in Chapter Five require imaging (e.g., MRI or SPECT) which is accessible in some but not all countries. These imaging tests are reasonably expensive tests that may restrict access for many people, but if they inform management that

aids recovery and return to work, they could potentially be cost-effective. Research into the diagnostic utility of these tests is important as if these tests do not guide care or provide prognostic information, then there is no use in investigating the issue of cost and availability of these index tests. Therefore, given there are now informative index tests the next step is to investigate the utility of these tests.

8.3.4. Implications for prognosis of non-specific low back pain

An important clinical use of a diagnosis is to provide patients with a more accurate prognosis. The results of Chapter Six demonstrated that some lumbar spine imaging findings (e.g., Modic changes, disc degeneration) were found to have some evidence of an association with future low back pain, however this was based on weak associations.[5] If these imaging findings are available to clinicians, they may assist clinicians to understand a patient's prognosis, however, it is currently not recommended to obtain imaging purely to gain this prognostic information. Currently, clinical guidelines recommend that clinicians should not refer patients for imaging for prognostic or diagnostic information.[24]

The results from the review in Chapter Six suggest imaging findings that are predictive of future low back pain differ depending on whether pain is present at baseline. For example, the presence of a high intensity zone was associated with better outcomes in those with low back pain but worse outcomes in those without low back pain at baseline. This is an important finding that may lead to better understanding of the relationship between imaging findings and low back pain. A recent cohort study (n=3369)[25] investigating the relationship between MRI findings and future low back pain, also found that spondylolisthesis was associated with better outcomes in those with low back pain at baseline, but worse future pain in those without low back pain at baseline. It is also possible that the relationship between some imaging findings

and future low back pain may be different in people with and without low back pain. Some MRI findings may have a different clinical profile and natural history compared to others, and research needs to account for these differences when conducting studies into the association between MRI findings and future low back pain.

8.3.5. Implications for stratified care and low back pain

In Chapter Five we investigated the importance of a diagnosis by investigating tests to identify the nociceptive source of low back pain and in Chapter Six explored how these specific pathoanatomical structures may potentially provide prognostic information. The aim of both chapters, to better understand whether a diagnosis would provide prognostic information or lead to more targeted treatments. However, another approach to determine the prognosis of patients with low back pain and potentially provide more targeted treatment is by exploring stratified care approaches (i.e., STarTBack Tool). The STarTBack Tool is an important method to stratify patients based on their risk level and provide matched care accordingly.[15] The results from Chapter Seven highlighted that there were large differences in some study level factors in relation to the implementation of the STarTBack approach, which may explain the differences in results between studies.

The original STarTBack RCT (Hill et al 2011[15]) had a more explanatory design while in many of the subsequent studies, the design was more pragmatic/real world. Pragmatic study designs portray the complex nature of clinical care in real life and the difficulties of implementing matched care. A key issue is that all studies in Chapter Seven sub-grouped participants based on their risk scores, however not all participants went on to receive matched interventions. If patients are not receiving the matched interventions, then the potential utility of the tool is not realised. For example, patients in medium or high-risk groups should receive treatments

inclusive of psychosocial interventions as per the tool. A recent randomised trial (n=492) found that cognitive functional therapy (CFT) was effective in reducing disability at follow-up (RMDQ: –4·6; 95% CI –5·9 to –3·4) in those with chronic low back pain, with most (90%) of the patients in the medium to high-risk group based on the STarTBack Tool. Therefore, providing the appropriate matched care as per the tool is arguably a vital task for clinicians. Chapter Seven highlighted that it was often not clearly reported whether participants received matched treatments (e.g., only 3 of 11 studies providing data on whether patients received matched care), limiting our ability to explore this further. The effectiveness of the STarTBack approach is also likely dependent on the amount of training of clinicians delivering the matched care, and therefore the clinical skills to provide the matched treatments. There was substantial variability in the amount of training provided to clinicians (range: 6 hours to 10 days) and in half the studies it wasn't clear the amount of training clinicians received or if training of clinicians was even provided.

The results of Chapter Seven were limited due to poor reporting of data in studies for some of the key study level factors. For example, reporting on whether participants received matched care is critical to interpreting the results of studies, as it likely contributes to the different outcomes in studies. Future research needs to better address the lack of reporting in how the STarTBack Tool, or other stratified approaches, are implemented. Better reporting in the literature will provide researchers a better understanding of the barriers to the effectiveness of stratified approaches and can guide future research to design explanatory study designs before conducting pragmatic studies.

8.4. Suggestions for future research

Future research has the challenge of addressing the key knowledge gaps highlighted in the chapters of this thesis. In Chapter Four, suggests a need to investigate the diagnostic accuracy of clusters of red flags rather than individual red flags. Chapter Five and Six and highlight the need to investigate the clinical utility of diagnostic tests and MRI findings respectively, the nuances of different thresholds used for different index tests and MRI findings, and the lack of standardisation of outcome measures and MRI diagnostic classifications/criteria in the field of low back pain. Chapter Seven demonstrates the need for future research to investigate how to best implement the STarTBack approach to address the differing results of the effectiveness of the STarTBack approach.

8.4.1. Future research for serious low back pain

The Cochrane review in Chapter 4 suggests that we need more research, and better research, evaluating red flags for vertebral fracture. Most of the studies included in the Cochrane review evaluated single red flags, however this is generally not how red flags are used in clinical practice. Future research should attempt to develop clinical decision rules based on clusters of findings, which has previously been done in other areas of the body (e.g., Ottawa Ankle Rules[26], Ottawa Knee Rules[27], Canadian C Spine Rules[28]). These clinical decision rules are sensitive and have been shown to reduce unnecessary imaging. The future research evaluating these clinical decision rules will provide a useful roadmap for a research program aiming to develop and validate a clinical decision rule for vertebral fracture in patients with low back pain. Future research investigating clinical decision rules for red flags will need to include rule validation and evaluation and investigate the downstream impact of applying such rules in routine care. Future research should also consider whether these clinical decision rules need to be different for the different types of vertebral fracture.

8.4.2. Future research for diagnosis in non-specific low back pain

The diagnostic systematic review in Chapter Five highlights now that a diagnosis within non-specific low back pain is possible, more research into the utility of these tests is needed. It should be a research priority to investigate whether patients judged to have low back pain due to the disc, sacroiliac joint, or facet joint based upon validated index tests have different prognoses or responses to treatment compared to patients considered not to have those conditions. There is also a need to investigate whether identifying a pathoanatomical source of low back pain can lead to better selection of existing treatments or even prevention of future episodes of low back pain. Future research needs to explore the pathological mechanisms leading to pathoanatomical changes. For example, understanding the mechanisms leading to Modic type 1 changes will create opportunities for new and more specific targeted treatments, and help researchers and clinicians better understand prognosis. This has been successful in other areas of medicine. A prime example is rheumatoid arthritis, once a debilitating condition leading to deformities and joint damage, advancements in understanding pathological mechanisms behind rheumatoid arthritis have led to the development of more targeted treatments resulting in better outcomes.

There are also concerns around optimal thresholds for a positive test when using continuous or ordinal index tests (e.g., MRI evidence of disc degeneration grade 3 compared to grade 5). Future research needs to explore whether differing degrees of severity of a condition e.g., disc degeneration grade 4-5 has a different prognosis or response to treatment than those with lower grades of degeneration (e.g., grade 1-2). An important consideration is the impact of age on different levels of severity, for example, a 20-year-old with disc degeneration grade 4-5 may have a different natural course than an 80-year-old with similar levels of disc degeneration. Future research in the low back pain field should follow the research performed in other areas

of medicine. For example, if a patient's blood glucose levels begin to rise, we understand that there is a threshold in which if those levels do not decrease, they will likely develop a disease (i.e., diabetes) which is also associated with other health comorbidities (e.g., high cholesterol, kidney disease etc).

A key implication from the findings in Chapter Five is that future research needs to clarify the clinical value of identifying pathoanatomical structures using diagnostic tests, as important research has shown unnecessary testing can lead to increases in resources, costs, and negative downstream consequences.[23] However, a greater understanding of the nociceptive, pathoanatomical, and/or psychosocial causes of low back pain may reduce over-imaging, overdiagnosis, and consequent overtreatment. Although diagnostic studies are challenging to conduct, they have the potential to reduce unnecessary imaging and its associated costs for low back pain, in a field that has been challenged with over-imaging.[23] Future high-quality randomised trials need to be performed to now explore whether the pathoanatomical structures identified using such diagnostic tests has the ability to identify patients who will benefit from further testing. The ultimate aim of this research is to improve patient outcomes in those unresponsive to first line care.

8.4.3. Future research for prognosis of non-specific low back pain

The relationship between a diagnosis (i.e., through clinical examination findings and/or imaging findings) and response to treatment needs to be further explored. For instance, identifying diagnostic tests with high diagnostic accuracy will open opportunities to better understand the association between a diagnosis (e.g., MRI evidence of Modic change) and whether a diagnosis provides prognostic information on the natural course of the condition or response to a specific treatment. Contrary to popular belief, there has been very little research

investigating if a diagnosis identifies patients who respond to a specific treatment.[29] For example, Steffens et al 2016[29] systematic review attempted to investigate if MRI findings predict patients who respond better to specific treatments. This review only included eight surgical or injection intervention trials and results were based on single studies with small sample sizes.[29] Future research needs to prioritise these issues through large scale high quality randomised trials or gain access to independent patient data (IPD) to perform IPD meta-analysis and explore interactions between diagnosis and treatment. For IPD meta-analysis to be possible, there will need to be standardisation in the literature of low back pain outcome measures and imaging findings (e.g., MRI findings).[30] Another study method to better understand responses to treatment can be through high-quality longitudinal studies aiming to determine the clinical relevance of imaging findings (e.g., MRI) and low back pain. Longitudinal studies will need to adjust for other key prognostic factors such as age, a patient's psychosocial profile, and severity of imaging findings (e.g., Pfirrman scale).

Our systematic review in Chapter Six highlighted that there is a multitude of existing data in the literature demonstrating potential associations between MRI findings and future low back pain; however, much of this data is difficult to utilise due to heterogeneity between studies (i.e., reporting of clinical outcomes, imaging findings, statistical measures). Therefore, standardising the assessment and reporting of clinical outcomes based on recommended core outcome measures for those with low back pain is important to utilise the existing data in the literature.[30] Standardising clinical outcome measures will allow for better comparisons between study results and improves the interpretability of results in this area. Future studies should aim for standardised imaging protocols and sequences. For example, MRI protocols should include standardised descriptions and thresholds for positive MRI findings. For example, for disc degeneration, researchers could employ a standardised threshold to classify

the severity of disc degeneration using the Pfirrman classification (grades 1 to 5), which has proven to be useful in research.[31] A recent exploratory, cross-sectional cohort study (n=412) found that the thresholds used to define the presence of lumbar disc degeneration influence how strongly it is associated with low back pain.[32] The Pfirrman classification has demonstrated good inter and intra-observer agreement[33] and may be an option moving forward to standardise thresholds in research. Lastly, future research should investigate the impact of age and psychosocial factors on associations between imaging findings and clinical outcomes.

8.4.4. Future research for stratified care and low back pain

Chapter Seven highlighted the conflicting results of studies investigating the effectiveness of the STarTBack approach. Given the results of Chapter Seven, it is important that future research evaluates stratified care (e.g., STarTBack approach) further. One approach is using an explanatory research design, like the original Hill et al 2011[15] study to determine if there is merit in continuing with subsequent pragmatic studies. Alternatively, another approach is that future research should explore implementation strategies to optimise the delivery of the STarTBack approach. Following the initial Hill et al 2011[15] study, very few studies that used a more pragmatic study design reported on the key challenges surrounding the implementation of the STarTBack approach. A better understanding of how to best implement the STarTBack approach to ensure greater compliance with the recommendations of the tool is needed, and important nuances on implementation methods and feasibility needs to be further investigated. Qualitative research to better understand the barriers clinicians face when using the STarTBack Tool in practice and develop strategies to address barriers to compliance, could be important to improving implementation. A better understanding of the key variables influencing the implementation of matched treatments may help inform clinicians of the best approach to implement stratified care in clinical practice. The STarTBack approach needs to be further

evaluated as currently there is not enough evidence in the literature to make definitive conclusions regarding its effectiveness in managing patients with low back pain.

8.5. Conclusion

The studies included in this thesis addressed key issues in diagnosis and targeted treatment for low back pain. Further high-quality research in this field is needed, as there are clear challenges in progressing diagnostic research in low back pain. However, the studies in this thesis provide researchers and clinicians a foundation to move the field forward. To improve care for patients with low back pain and reduce low value care (e.g., overdiagnosis), research efforts need to be made to identify those patients where a diagnosis can provide prognostic information, potentially prevent low back pain, and/or improve patient outcomes through more targeted and effective treatment approaches.

8.6. References

1. Han CS, Hancock MJ, Maher CG. Reconsidering non-specific low back pain: where to from here? *Spine J.* 2022;22(12):1927-1930.
2. Shaw EG, Taylor JG. The results of lumbo-sacral fusion for low back pain. *J Bone Joint Surg Br.* 1956;38-B(2):485-497.
3. Waddell G. 1987 Volvo award in clinical sciences. A new clinical model for the treatment of low-back pain. *Spine (Phila Pa 1976).* 1987;12(7):632-644.
4. Han CS, Hancock MJ, Sharma S, et al. Low back pain of disc, sacroiliac joint, or facet joint origin: a diagnostic accuracy systematic review. *eClinicalMedicine.* 2023;59:101960.
5. Han CS, Maher CG, Steffens D, et al. Some magnetic resonance imaging findings may predict future low back pain and disability: a systematic review. *J Physiother.* 2023.
6. Beneciuk JM, George SZ. Pragmatic Implementation of a Stratified Primary Care Model for Low Back Pain Management in Outpatient Physical Therapy Settings: Two-Phase, Sequential Preliminary Study. *Phys Ther.* 2015;95(8):1120-1134.
7. Choudhry NK, Fifer S, Fontanet CP, et al. Effect of a Biopsychosocial Intervention or Postural Therapy on Disability and Health Care Spending Among Patients With Acute and Subacute Spine Pain: The SPINE CARE Randomized Clinical Trial. *JAMA.* 2022;328(23):2334-2344.
8. Delitto A, Patterson CG, Stevans JM, et al. Stratified care to prevent chronic low back pain in high-risk patients: The TARGET trial. A multi-site pragmatic cluster randomized trial. *EClinicalMedicine.* 2021;34:100795.
9. Konstantinou K, Lewis M, Dunn KM, et al. Stratified care versus usual care for management of patients presenting with sciatica in primary care (SCOPiC): a randomised controlled trial. *Lancet Rheumatol.* 2020;2(7):e401-e411.
10. Koppenaal T, Pisters MF, Kloek CJ, Arensman RM, Ostelo RW, Veenhof C. The 3-Month Effectiveness of a Stratified Blended Physiotherapy Intervention in Patients With Nonspecific Low Back Pain: Cluster Randomized Controlled Trial. *J Med Internet Res.* 2022;24(2):e31675.
11. Morso L, Olsen Rose K, Schiottz-Christensen B, Sowden G, Sondergaard J, Christiansen DH. Effectiveness of stratified treatment for back pain in Danish primary care: A randomized controlled trial. *Eur J Pain.* 2021;25(9):2020-2038.
12. Murphy SE, Blake C, Power CK, Fullen BM. Comparison of a Stratified Group Intervention (STarT Back) With Usual Group Care in Patients With Low Back Pain: A Nonrandomized Controlled Trial. *Spine (Phila Pa 1976).* 2016;41(8):645-652.
13. Rhon DI, Greenlee TA, Poehlein E, et al. Effect of Risk-Stratified Care on Disability Among Adults With Low Back Pain Treated in the Military Health System: A Randomized Clinical Trial. *JAMA Netw Open.* 2023;6(7):e2321929.
14. Cherkin D, Balderson B, Wellman R, et al. Effect of Low Back Pain Risk-Stratification Strategy on Patient Outcomes and Care Processes: the MATCH Randomized Trial in Primary Care. *J Gen Intern Med.* 2018;33(8):1324-1336.
15. Hill JC, Whitehurst DG, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet.* 2011;378(9802):1560-1571.

16. Foster NE, Mullis R, Hill JC, et al. Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. *Ann Fam Med.* 2014;12(2):102-111.
17. Han CS, Hancock MJ, Downie A, et al. Red flags to screen for vertebral fracture in patients presenting with low back pain. *Cochrane Database of Systematic Reviews.* 2022(7).
18. Verhagen AP, Downie A, Popal N, Maher C, Koes BW. Red flags presented in current low back pain guidelines: a review. *Eur Spine J.* 2016;25(9):2788-2802.
19. Sugiura S, Aoki Y, Toyooka T, et al. Utility of Physical Examination Findings for Predicting Low-Back Pain in Adolescent Patients with Early-Stage Spondylolysis: A Retrospective Comparative Cohort Study. *Spine Surg Relat Res.* 2021;5(6):412-417.
20. Therriault T, Rospert A, Selhorst M, Fischer A. Development of a preliminary multivariable diagnostic prediction model for identifying active spondylolysis in young athletes with low back pain. *Phys Ther Sport.* 2020;45:1-6.
21. Saueressig T, Owen PJ, Diemer F, Zebisch J, Belavy DL. Diagnostic Accuracy of Clusters of Pain Provocation Tests for Detecting Sacroiliac Joint Pain: Systematic Review With Meta-analysis. *J Orthop Sports Phys Ther.* 2021;51(9):422-431.
22. O'Keeffe M, Ferreira GE, Harris IA, et al. Effect of diagnostic labelling on management intentions for non-specific low back pain: A randomized scenario-based experiment. *Eur J Pain.* 2022;26(7):1532-1545.
23. Jacobs JC, Jarvik JG, Chou R, et al. Observational Study of the Downstream Consequences of Inappropriate MRI of the Lumbar Spine. *J Gen Intern Med.* 2020;35(12):3605-3612.
24. Oliveira CB, Maher CG, Pinto RZ, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *Eur Spine J.* 2018;27(11):2791-2803.
25. Kasch R, Truthmann J, Hancock MJ, et al. Association of Lumbar MRI Findings with Current and Future Back Pain in a Population-based Cohort Study. *Spine (Phila Pa 1976).* 2022;47(3):201-211.
26. Jonckheer P, Willems T, De Ridder R, et al. Evaluating fracture risk in acute ankle sprains: Any news since the Ottawa Ankle Rules? A systematic review. *Eur J Gen Pract.* 2016;22(1):31-41.
27. Sims JI, Chau MT, Davies JR. Diagnostic accuracy of the Ottawa Knee Rule in adult acute knee injuries: a systematic review and meta-analysis. *Eur Radiol.* 2020;30(8):4438-4446.
28. Michaleff ZA, Maher CG, Verhagen AP, Rebbeck T, Lin CW. Accuracy of the Canadian C-spine rule and NEXUS to screen for clinically important cervical spine injury in patients following blunt trauma: a systematic review. *CMAJ.* 2012;184(16):E867-876.
29. Steffens D, Hancock MJ, Pereira LS, Kent PM, Latimer J, Maher CG. Do MRI findings identify patients with low back pain or sciatica who respond better to particular interventions? A systematic review. *Eur Spine J.* 2016;25(4):1170-1187.
30. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain.* 2018;159(3):481-495.
31. Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine (Phila Pa 1976).* 2001;26(17):1873-1878.

32. Dragsbaek L, Kjaer P, Hancock M, Jensen TS. An exploratory study of different definitions and thresholds for lumbar disc degeneration assessed by MRI and their associations with low back pain using data from a cohort study of a general population. *BMC Musculoskelet Disord.* 2020;21(1):253.
33. Urrutia J, Besa P, Campos M, et al. The Pfirrmann classification of lumbar intervertebral disc degeneration: an independent inter- and intra-observer agreement assessment. *Eur Spine J.* 2016;25(9):2728-2733.