Received: 30 August 2023 | Accepted: 31 August 2023

DOI: 10.1002/msc.1816

SYSTEMATIC REVIEW

WILEY

Diagnostic clinical prediction rules for categorising low back pain: A systematic review

Charles James Hill | Anirban Banerjee | Jonathan Hill | Claire Stapleton

University of Keele, Newcastle under Lyme, UK

Correspondence
Jonathan Hill.
Email: j.hill@keele.ac.uk

Abstract

Background: Low back pain (LBP) is a common complex condition, where specific diagnoses are hard to identify. Diagnostic clinical prediction rules (CPRs) are known to improve clinical decision-making. A review of LBP diagnostic-CPRs by Haskins et al. (2015) identified six diagnostic-CPRs in derivation phases of development, with one tool ready for implementation. Recent progress on these tools is unknown. Therefore, this review aimed to investigate developments in LBP diagnostic-CPRs and evaluate their readiness for implementation.

Methods: A systematic review was performed on five databases (Medline, Amed, Cochrane Library, PsycInfo, and CINAHL) combined with hand-searching and citation-tracking to identify eligible studies. Study and tool quality were appraised for risk of bias (Quality Assessment of Diagnostic Accuracy Studies-2), methodological quality (checklist using accepted CPR methodological standards), and CPR tool appraisal (GRade and ASsess Predictive).

Results: Of 5021 studies screened, 11 diagnostic-CPRs were identified. Of the six previously known, three have been externally validated but not yet undergone impact analysis. Five new tools have been identified since Haskin et al. (2015); all are still in derivation stages. The most validated diagnostic-CPRs include the Lumbar-Spinal-Stenosis-Self-Administered-Self-Reported-History-Questionnaire and Diagnosis-Support-Tool-to-Identify-Lumbar-Spinal-Stenosis, and the StEP-tool which differentiates radicular from axial-LBP.

Conclusions: This updated review of LBP diagnostic CPRs found five new tools, all in the early stages of development. Three previously known tools have now been externally validated but should be used with caution until impact evaluation studies are undertaken. Future funding should focus on externally validating and assessing the impact of existing CPRs on clinical decision-making.

KEYWORDS
clinical decision support system, clinical prediction rules, decision support tool, diagnostic accuracy, diagnostic rules, low back pain, predictive models

1 | INTRODUCTION

Low back pain (LBP) is the single largest cause of long-term disability in England (Institute for Health Metrics and Evaluation, 2020), with a growing prevalence in the United Kingdom (UK) and other developed countries, particularly in ageing populations (Hartvigsen et al., 2018). As such, LBP represents a major public health issue with significant consequences for individuals, society, and the economy. The impact of LBP on healthcare systems is substantial, accounting for a significant proportion of healthcare expenditures, with back pain alone accounting for the highest overall disability burden from all diseases in the UK, at 11% of total spend within the National Health Service (NHS England National Pathfinder Project, 2014).

Among health professionals, physiotherapists spend a significant amount of time with patients and play a major role in the rehabilitation of patients with LBP (Foster et al., 2018). However, the effectiveness of physiotherapy treatments designed to improve LBP remains suboptimal, which is partly attributed to a tendency to treat nonspecific LBP as one homogenous condition (Herbert et al., 2011). The term nonspecific LBP describes many different subgroups of conditions which may respond preferentially to specific treatments (Saragiotto et al., 2017). Some specific LBP conditions have traditionally been difficult to identify as diagnosis is based on expert opinion and biological plausibility, with poor concordance found between clinicians (Foster et al., 2018).

Certain conditions such as lumbar spinal stenosis (LSS) and axial spondyloarthritis (axSpA) can be difficult to identify as there is no reliable reference standard criteria for diagnosis (Jensen et al., 2020). In LSS, symptoms can be similar to other lower back conditions, such as herniated discs or sciatica, and diagnostic imaging such as MRI or CT scans may not always provide clear evidence of the condition (Cook et al., 2020). In some cases, patients may undergo unnecessary surgery or other invasive procedures because of an inaccurate diagnosis (Jensen et al., 2020). To address this challenge, recent research has focused on developing more accurate diagnostic tools with cut-off points to rule in or rule out specific conditions. Within the LBP literature, these are often termed 'clinical prediction rules' (CPRs). CPRs consider both the literature and expert opinion to produce algorithms based on quantitative prediction models (Beattie & Nelson, 2006). Generally, they are applied to support decision-making in three areas: (1) diagnosis, (2) prognosis, or (3) treatment response (Herbert et al., 2011). They have been shown to have several benefits, including reducing the need for unnecessary imaging, improving the accuracy of clinical assessment, and enabling more timely initiation of treatment (Scott & Crock, 2020).

Diagnostic CPRs are a type of CPR that uses an algorithm to estimate the probability of a specific diagnosis based on a patient's assessment features. Given the complexity of the diagnostic process, any algorithm used for diagnosis must include variables with high accuracy (sensitivity and specificity) to successfully detect the target condition (Cook et al., 2020). To identify these variables, clinicians typically collect assessment and demographic data on a cohort of patients and then use logistic regression analysis to explore the association of diagnostic variables with the condition of interest (Cook, 2008). Through this process, the most predictive items are retained within the CPR in order to predict the probability of a diagnosis (Cook et al., 2020). After its initial development, the diagnostic CPR should undergo an external validation process whereby it is used on diverse patient groups and in diverse settings to assess its capability to predict the same diagnosis with accuracy. Once a diagnostic CPR is validated, it can be clinically implemented (Khalifa et al., 2019), and an impact analysis should be performed to evaluate its effectiveness in clinical decision-making, enhancing patient outcomes, or optimising resource utilisation (van Geloven et al., 2022).

The development stage of a diagnostic CPR is believed to be crucial for determining its suitability for use in clinical practice. If a diagnostic CPR has not undergone external validation, it should not be used in practice as it may only reflect chance statistical associations or be specific to the patient sample or setting from which it was developed (McGinn et al., 2008). To determine the suitability of a diagnostic CPR for use in similar patient populations, it is necessary to validate its diagnostic accuracy in new (external to the derivation sample) patient cohorts across different clinical settings. However, it is important to note that even a validated diagnostic CPR may not necessarily be more accurate than unassisted clinician judgement, and its application may not always result in beneficial clinical outcomes (Steyerberg, 2009). Therefore, impact analysis is required to determine the potential benefits of applying a diagnostic CPR in clinical practice with confidence (McGinn et al., 2008).

Numerous systematic reviews have investigated the effectiveness of CPRs to support decision making in LBP treatment (Beneciuk et al., 2009; May & Rosedale, 2009; Stanton et al., 2010; van Oort et al., 2012; Haskins et al., 2012, 2015; Patel et al., 2013; Lubetzky-Vilnai et al., 2014; Peterson et al., 2017; Cook et al., 2020). These reviews focus primarily on the use of CPRs for predicting treatment outcomes or prognosis. It has been found that most prognostic CPRs are still in the initial derivation phases of development and have not yet undergone external validation or a comprehensive impact analysis (Binuya et al., 2022; Haskins et al., 2015). Only two systematic reviews, Haskins et al. (2012, 2015), have specifically examined diagnostic CPRs, which diagnose subtypes of LBP. Consequently, little is known about any diagnostic CPRs which have been developed post-2015, or if any of the tools identified in this previous systematic review have since undergone further validation.

Therefore, the aims of this study were to (1) build on the findings of Haskins et al. (2015) and investigate any developments in diagnostic CPR research since its publication and (2) to summarise the evidence of existing LBP diagnostic CPRs, and evaluate their readiness for clinical practice.

2 | METHODS

A systematic review was selected as this is the gold standard method for synthesising evidence from the literature (Munn et al., 2018). To ensure transparent and comprehensive reporting of the review

findings, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines 2020 were followed (Page et al., 2021).

The full operational description of diagnostic CPRs in this review is described in Table 1, based on descriptions of diagnostic decision-making tools from Aggarwal et al. (2015) and Haskins et al. (2015).

2.1 | Eligibility criteria

The review's eligibility criteria (Table 2) were adapted from published protocols of previous reviews in this area (Haskins et al., 2015; Van Oort et al., 2012).

2.2 | Search strategy

To identify relevant diagnostic CPRs for the nonsurgical management of adults with LBP, a thorough two stage systematic literature search was conducted to capture derivation, validation, or impact analysis studies. Firstly, five databases were searched (Medline, AMED, Cochrane Library, PsychInfo, and CINAHL) using highly sensitive adaptations of search strings proposed by Geersing et al. (2012) for identification of diagnostic studies. Studies were captured from January 2013 to January 2023 to capture the literature since the Haskins et al. (2015) review. Where appropriate, the strategy contained Boolean operators, truncations, and MeSH headings. The full search strategy (Appendix A1) was consistent with terms used previously in similar systematic reviews to identify CPRs related to prognosis and diagnosis (Beneciuk et al., 2009; Haskins et al., 2012, 2015; Van Oort et al., 2012). It was concluded that the previous systematic review by Haskins et al. (2015) captured all relevant literature from the inception to 2013 for all five databases. Studies identified in the search were uploaded to the reference manager software RefWorks for duplicate removal. Secondly, hand searching and citation tracking were later used as supplementary search strategies for identified studies. Table 3 illustrates one of the five database search strategies used.

The search results were processed using Rayyan, a screening software developed by Ouzzani et al. (2016) (https://rayyan.ai), after removing duplicates using RefWorks. A reviewer (CH) assessed the title and abstract eligibility of identified studies, with a second blinded reviewer (PG) screening a random 10% of titles and abstracts. The agreement between the two reviewers was determined using a Cohen's Kappa coefficient (κ). Kappa was categorised according to Landis and Koch (1977) with values of 0–0.2 indicating 'slight', 0.21–0.40 'fair', 0.41–0.60 'moderate', 0.61–0.80 'substantial', and 0.81–1.0 'almost perfect'. Potentially eligible studies underwent full-text screening (CH), with eligible studies listed for full data extraction. A further random set of full-text studies (10%) underwent secondary reviewer blinded screening (PG) and the κ agreement between reviewers was calculated. In the event of disagreement between the reviewers, a consensus meeting was held with a third independent reviewer (CS) to provide the final judgement.

TABLE 1 Operational features of diagnostic clinical prediction rules (CPRs) included in this review.

The diagnostic CPR is a clinical decision-making tool using patient data to estimate the likelihood of a specific diagnosis.
The diagnostic CPR may use a combination of clinical (physical examination, imaging, laboratory tests), demographic (patient's age, sex, race, ethnicity), and environmental data (patient's occupation or lifestyle) to generate a risk score or probability estimate for a LBP condition.
The goal of the diagnostic CPR should be to improve diagnostic accuracy and identify patients who may benefit from further diagnostic testing or referral to a specialist.
The diagnostic CPR should be developed and validated using large datasets and formal statistical methods to ensure that it is accurate and reliable for detecting a diagnosis.
The diagnostic CPR is distinct from classification criteria and screening tools, which are primarily used to identify patients with certain features or risk factors of disease who need further testing, but do not provide a specific probability estimate for diagnosis of a condition.

TABLE 2 Study eligibility criteria.

Inclusion criteria
1. Studies reporting on the derivation, validation, or impact analysis of a diagnostic CPR related to the nonsurgical management of adults with LBP
2. The tool under development contains 2 or more predictor variables
3. The tool was derived using a formal statistical method such as logistic regression analysis whereby candidate predictor variables are selected for inclusion in the final diagnostic CPR due to their association with specific LBP conditions
4. The tool is presented in sufficient detail as to inform clinical diagnosis
Exclusion criteria
1. Ineligible article types: Conference proceedings/abstracts, dissertations, commentaries, reviews, editorials, letters, study protocols, case reports, books, book reviews, clinical practice guidelines
2. Derivation studies of screening tools and classification criteria
3. Studies published before 2013; except for those detailed in the Haskins et al. (2015) paper
4. Studies not written in the English language
5. Studies with no full text available

Note: (1) Non-surgical management of adults was chosen to limit irrelevant tools around. (2) Studies were chosen from 2013 onwards as the search strategy of Haskins et al. (2015) dated from database inception to 2013. (3) As Haskins et al. (2015) was the latest relevant systematic review, the author concluded that this paper had accurately identified all diagnostic CPRs up to this point.

TABLE 3 CINAHL via EBESCO search strategy.

S5	S1 and S4 and (S2 OR S3) limiters: English language, excluding MEDLINE records, humans only, full text only, 2013–2023 (n = 393)
S4	(Diagnosis or diagnosing or diagnostics)
S3	Stratification OR mh 'ROC Curve' OR discrimination OR discriminate OR c-statistic OR c statistic OR 'Area under the curve' OR AUC OR calibration OR indices OR algorithm OR multivariable
S2	(Validat* OR ti predict* OR rule*) OR (predict* AND (outcome* OR risk* OR model*)) OR ((history OR variable* OR criteria OR scor* OR characteristic* OR finding* OR factor*) AND (predict* OR model* OR decision* OR identif* OR prognos*)) OR (decision* AND (model* OR clinical* OR MH 'logistic regression+')) OR (prognostic AND (history OR variable* OR criteria OR scor* OR characteristic* OR finding* OR factor* OR model*))
S1	'Dorsalgia' OR (MH 'Back Pain+') OR (MH 'Low Back Pain') OR 'backache' OR (lumbar W1 pain) OR (lumbar N5 pain) OR (MH 'Coccyx') OR (MH 'Sciatica') OR 'sciatica' OR 'coccyx' OR 'coccydynia' OR 'back disorder*' OR (MH 'Lumbar Vertebrae') OR (lumbar N2 vertebra) OR (MH 'Thoracic Vertebrae') OR (MH 'Spondylolisthesis') OR (MH 'Spondylolysis') OR 'lumbago'

2.3 | Data extraction and quality appraisal

Risk of bias (RoB) was evaluated using the validated QUADAS-2 tool (Quality Assessment of Diagnostic Accuracy Studies-2) designed for multivariable prediction studies (Whiting et al., 2011). Following the approach described by McGinn et al. (2000), all subsequent derivation and validation studies for each diagnostic CPR identified were also considered for evaluation. An additional standardised methodological appraisal of studies included for analysis was performed in keeping with the approach of two previous systematic reviews of CPRs (Haskins et al., 2012, 2015). Each criterion was scored as 'high', 'low', or 'unclear' RoB. Full tables can be found in Appendix A2. This tool was specifically selected because it contains items that follow international standards for CPR methodological development (Beattie & Nelson, 2006; Steel et al., 2012). The QUADAS-2 and the methodological quality appraisal tool were independently applied by one reviewer (CH).

The validated Grade and Assess Predictive tool (GRASP) framework was used to quality appraise the diagnostic CPRs identified (Khalifa et al., 2019). The GRASP tool was chosen because unlike previous appraisal approaches such as the 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis' (TRIPOD) statement (Collins et al., 2015), or the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist (Moons et al., 2014), it evaluates the predictive performance of CPRs as well as usability and post-implementation impact (Khalifa et al., 2019). Although the GRASP framework is not specific to diagnostic CPRs but is relevant to prediction tools more generally, it was still appropriate to use this framework when appraising diagnostic CPRs identified (Khalifa et al., 2019).

For each study, data were extracted into a Microsoft Excel (2023) document to reflect the GRASP tool items: diagnostic CPR name, authors, year, intended use, intended user, category (diagnostic, prognostic or treatment responsive), clinical area, target population, target outcome, action, input source, input type, local context, study methodology (design, participant characteristics, reference standard, statistical methods, outline of the tool itself, cut-off points for diagnosis, performance) (sensitivity, specificity, calibration, area under ROC curve [AUC]), endorsement, automation flag (manual or automatic), total tool citations, total studies reported in, the phase of evaluation, grade assigned, direction of current evidence, and justification for the assigned grade (see Khalifa et al., 2019). The calibration in predictive modelling refers to the degree of agreement between the predicted probabilities of an event occurring and the actual proportion of observed events (Van Calster et al., 2023).

3 | RESULTS

3.1 | Study selection

The search strategy yielded 5021 studies, which following duplicate removal (n = 329) and the addition of studies identified from citation searching (n = 33), resulted in 4725 unique studies for title and abstract screening. Following screening, there were 160 studies sought for full text retrieval, of which three were not available in English. A total of 16 studies met the inclusion/exclusion criteria and reported either the derivation, validation, or impact analysis of a diagnostic CPR. See Figure 1 for the full PRISMA flow diagram.

Agreement between reviewers (CH and PG) was 'almost perfect' (κ = 0.97; 95% CI: 0.95, 0.99; absolute agreement, 98.3%) for the 10% of titles and abstracts screened (n = 474) and 'almost perfect' (κ = 0.80; 95% CI: 0.48, 1.13; absolute agreement, 86.7%) for the 10% of full text articles screened (n = 15). Disagreements during the full text stage of screening (n = 3) were resolved by consensus without the need for a third reviewer (CS). See Appendix A4 for full calculations.

3.2 | Study characteristics

The 16 studies included for analysis contained 11 separate LBP diagnostic CPRs created between 2007 and 2022. The review identified nine derivation studies, five validation studies, and two studies describing both derivation and validation. One study (Tominaga et al., 2022) described the diagnostic accuracy of two separate diagnostic CPRs (Konno, Kikuchi, et al. (2007), Konno, Hayashino, et al. (2007)). No impact analysis studies were found. Diagnostic CPRs were developed in Japan (n = 3), Germany (n = 2), France (n = 1), the UK (n = 1), United States (n = 1), Thailand (n = 1), China (n = 1) and one was developed in the United States and later validated in the UK. The characteristics of identified studies are

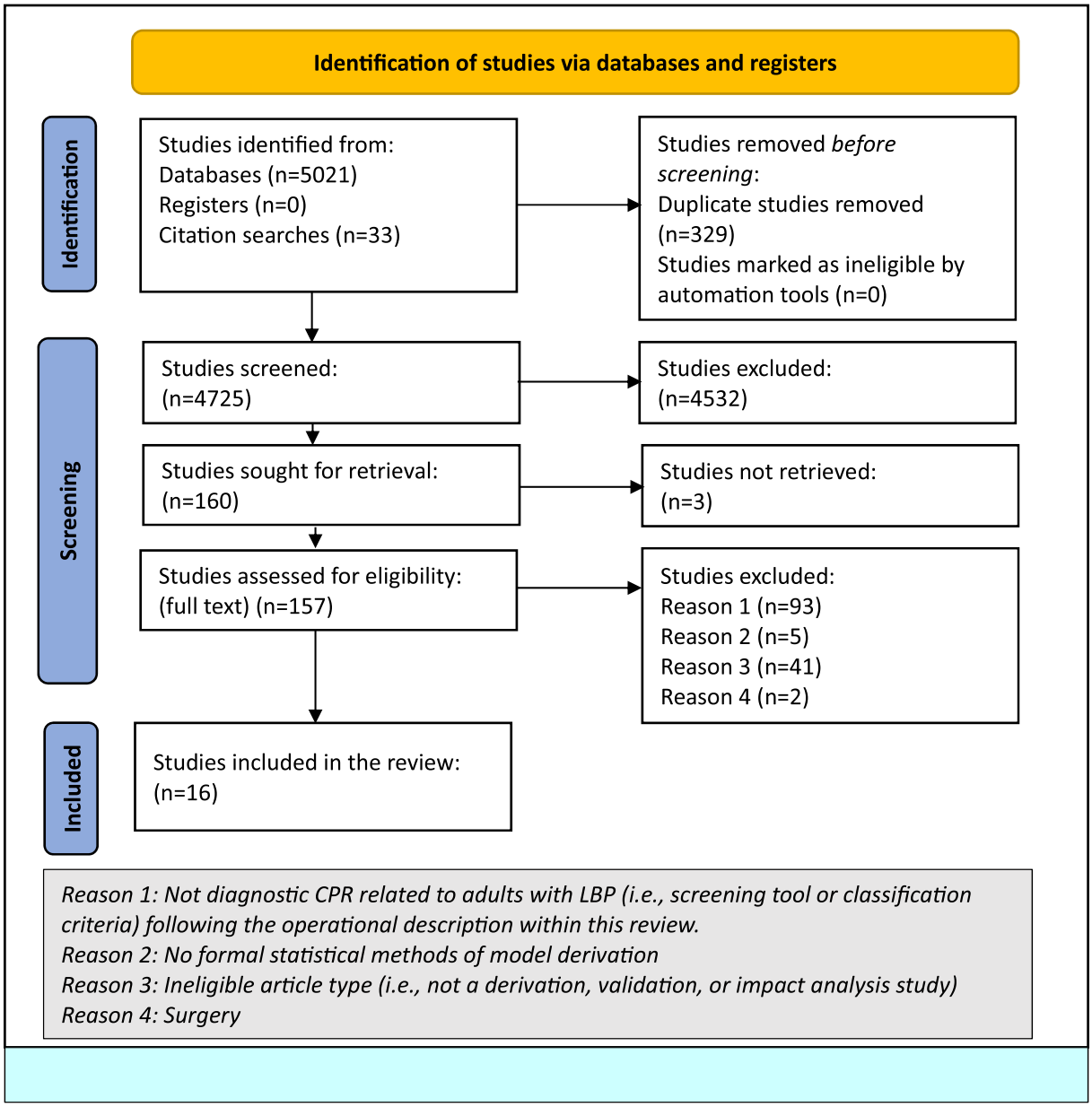

FIGURE 1 PRISMA 2020 reporting of study screening process.

described in Table 4 with additional detail provided in the GRASP forms (see Appendix A3). Published studies ranged in sample size from 86 to 33,545 patients and had a mean age between 36 and 70.5 years.

Of the 11 diagnostic CPRs found, diagnostic decision support was available for lumbar spinal stenosis (LSS) (n = 1), radicular or neurogenic claudication type LSS (n = 1), lumbar vertebral fracture (n = 1), osteoporotic vertebral compression fracture (OVCF) (n = 1), axial or radicular LBP (n = 1), axial spondyloarthritis (n = 2), sciatica (n = 1), lumbar instability (n = 1), multiple LBP diagnoses (n = 1) and occupation related LBP (n = 1). To indicate the likelihood of a diagnosis, diagnostic CPRs either used a risk scoring index (n = 9) or a classification and regression tree algorithm (n = 2). Of the 11 derivation studies, five employed a prospective design, four were retrospective, and two were cross-sectional.

3.3 | Methodological appraisal

Tables 5 and 6 present the methodological quality of identified derivation and validation studies using an approach previously seen in other CPR related systematic reviews. Among the 11 derivation studies (Table 5), aspects of quality that were poorly reported included predictor collinearity (n = 3/11), uncertainty in post-test probability (n = 2/11), and uncertainty in diagnostic CPR accuracy (n = 3/11). Except for Benditz et al. (2019), all studies provided a description of the mathematical techniques used in the derivation of the diagnostic CPR and important patient study characteristics. The four diagnostic CPR studies that had the highest methodological quality were Konno, Kikuchi, et al., 2007; Scholz et al., 2009; Stynes et al., 2018; Chatprem et al., 2021. The remaining seven studies met less than 50% of the appraisal items.

TABLE 4 Characteristics of included studies.

Tool name and related studies (first author)	No. of patients	Mean age (years)	Sex (male %)	Population studied	Condition of interest	Reference standard
LSS-DST						
Derivation Konno, Hayashino, et al. (2007)	468	65.0	45.9	Adults with numbness and pain in the legs	Lumbar spinal stenosis (LSS)	Unblinded consensus from expert panel of surgeons considering history, physical examination, and radiographic findings.
Validation Kato et al. (2009)	118	68.2	52.5	Adults with numbness and pain in the legs		Unblinded consensus from expert panel of surgeons
Tominaga et al. (2022)	3331	70.5	52.6	Adults with numbness and pain in the legs		Unblinded diagnosis by single orthopaedic physician
Vert frac						
Derivation Roux et al. (2007)	410	74.3	0	Females with osteoporosis aged 65–85 years with back pain	Vertebral fracture (vert frac)	Spinal radiograph showing a grade ≥1 vertebral fracture
LSS-SSHQ[a]						
Derivation Konno, Hayashino, et al. (2007)	115	69.5	47.2	Patients recovering from surgery with identified radicular or neurogenic claudication type LSS.	LSS: Radicular or neurogenic claudication type	Consensus from expert panel of surgeons
Validation Konno, Hayashino, et al. (2007)	250	59.5	49.0	-		-
Kato et al. (2015)	33,545	68.5	44.8	Adults with pain and numbness in the legs		Blinded consensus from expert panel of surgeons
Aghaei et al. (2015)	235	59.4	41.7	Consecutive adults >50 years of age		Blinded orthopaedic physician.
Tominaga et al. (2022)	3331	70.5	52.6	Adults with pain and numbness in the legs		Consensus from expert panel of surgeons (blinding unclear)
StEP						
Derivation Scholz et al. (2009)	130	57.5	49.0	Patients with diabetic polyneuropathy, postherpetic neuralgia or chronic LBP symptoms.	To differentiate between pain subtypes	Diagnosis by neurosurgeon, rheumatologist, and spinal physiotherapist following examination and imaging
Validation Scholz et al. (2009)	194	50.3	57.2	Adults with chronic LBP	Axial or radicular LBP	
OVCF						
Derivation Roman et al. (2010)	1448	56.5	40.5	Adults with low back pain with or without leg pain	Osteoporotic vertebral compression fracture (OVCF)	Radiographic findings interpreted by expert clinician
AxSpA						
Derivation Braun et al. (2011)	322	36.0	49.9	Chronic back pain >2 months but <10 years, aged between 16 and 45	Axial spondyloarthritis (AxSpA)	Expert opinion from rheumatologist

TABLE 4 (Continued)

Tool name and related studies (first author)	No. of patients	Mean age (years)	Sex (male %)	Population studied	Condition of interest	Reference standard
Keele SCIATICA						
Derivation Stynes et al. (2018)	394	49.8	40.0	Adults with low back related leg pain	Sciatica	Clinical diagnosis ± MRI findings
Regensburg						
Derivation Benditz et al. (2019)	111	59.0	47.7	German adults with back pain	Ankylosing spondylitis, facet joint arthritis, herniated disc, spondylodiscitis, osteoporotic vertebral fracture, lumbar spinal stenosis, and spondylolisthesis	
Validation Benditz et al. (2021)	86	49.0	53.5	German adults with back pain		Experienced spinal surgeon
Lx inst						
Derivation Chatprem et al. (2021)	140	36.0	38.6	Adults between 20 and 60 years with chronic low back pain (>3 months)	Lumbar instability (lx inst)	X-ray imaging read by a blinded trained observer
Occ LBP derivation						
Saengdao et al. (2021)	220	38.5	5	Adults with LBP by NMQ screening (nordic MSK questionnaire)	Occupational-LBP (occ LBP)	Diagnosis by three occupational medicine physicians
Clinical nomogram						
Derivation Ye et al. (2022)	638	38.1	66.1	Adults confirmed with either axSpA (n = 424) or non-axSpA (n = 214)	Axial spondyloarthritis	Rheumatologist confirmation using ASAS[b] criteria

[a]Self-Administered, Self-Reported History Questionnaire.
[b]Assessment of SpondyloArthritis international Society.

Methodological quality was considered high in three out of four validation studies (Kato et al., 2009, 2015; Tominaga et al., 2022). All but one study (Benditz et al., 2021) validated the tool externally on an independent sample in a different clinical setting and described the uncertainty in tool accuracy. All included validation studies accurately applied the diagnostic CPR in practice.

3.4 | Risk of bias in studies

Table 7 presents the RoB of identified studies (n = 16) using the QUADAS-2 tool. Five studies had a low overall RoB (Aghaei et al., 2015; Benditz et al., 2019, 2021; Kato et al., 2015; Roux et al., 2007; Ye et al., 2022) and nine studies demonstrated acceptable applicability (Aghaei et al., 2015; Benditz et al., 2019, 2021; Chatprem et al., 2021; Kato et al., 2009, 2015; Roux et al., 2007; Tominaga et al., 2022; Ye et al., 2022).

RoB concerning the 'patient selection' domain was generally low (n = 12) but considered high in four studies (Chatprem et al., 2021; Roman et al., 2010; Saengdao et al., 2021; Stynes et al., 2018). The RoB regarding the 'index test' domain was frequently unclear (n = 7) but was considered high in two cases (Scholz et al., 2009; Stynes et al., 2018). The RoB concerning the 'reference standard' domain was deemed high in five studies (Braun et al., 2011; Kato et al., 2009; Konno, Hayashino, et al., 2007; Saengdao et al., 2021; Tominaga et al., 2022), and unclear in three studies (Aghaei et al., 2015; Konno, Hayashino, et al., 2007; Scholz et al., 2009). The RoB concerning the 'flow and timings' domain was generally low (n = 14), but high in two studies (Kato et al., 2009; Konno, Hayashino, et al., 2007). Applicability concerns for studies were generally consistent with their respective RoB values, but there were fewer applicability concerns regarding the reference standard for clinical diagnosis.

Figure 2 illustrates the QUADAS-2 results of the 16 included studies. Limitations with respect to having a diagnostic reference standard led to the most frequent potential RoB among studies

TABLE 5 Derivation studies—methodological appraisal.

Study (first author)	Items[a]																			Total (Y)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
Konno, Kikuchi, et al. (2007)	Y	Y	N	Y	Y	Y	N	Y	Y	N	Y	Y	Y	Y	N	NA	N	N	Y	12
Roux et al. (2007)	Y	P	N	P	Y	P	Y	N	P	N	Y	P	Y	N	N	N	NA	N	Y	6
Konno, Hayashino, et al. (2007)	Y	P	N	Y	Y	N	Y	N	N	N	Y	P	Y	N	N	NA	N	N	Y	7
Scholz et al. (2009)	Y	Y	Y	N	Y	Y	Y	N	N	N	Y	Y	N	N	N	NA	Y	Y	Y	11
Roman et al. (2010)	N	Y	P	Y	Y	Y	N	N	P	N	Y	N	N	N	N	N	Y	N	N	6
Braun et al. (2011)	Y	P	N	Y	Y	N	Y	N	N	N	Y	P	Y	Y	N	NA	N	NA	Y	8
Stynes et al. (2018)	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	P	Y	N	N	NA	Y	Y	Y	13
Benditz et al. (2019)	N	Y	N	N	N	N	Y	N	Y	Y	N	N	N	N	N	N	N	N	Y	5
Saengdao et al. (2021)	N	Y	N	Y	Y	Y	N	N	P	N	Y	N	N	N	N	N	N	N	Y	6
Chatprem et al. (2021)	N	Y	Y	P	Y	Y	Y	Y	Y	Y	Y	P	Y	Y	Y	N	N	N	Y	13
Ye et al. (2022)	N	N	P	Y	Y	Y	N	Y	Y	Y	Y	N	Y	N	N	N	N	N	Y	9

Note: The collinearity of predictors refers to variables in the model which are highly correlated with each other, so the unique contribution of each predictor on the outcome variable is difficult to determine. By not establishing collinearity of predictors this can lead to overfitting, where the model becomes too complex and is not generalisable to new patient groups. The post-test probability describes the probability of the patient having a condition after considering the result of the tool. Therefore, describing the uncertainty of this measure can help clinicians determine the confidence in the tools prediction (i.e. through confidence intervals), to account for uncertainty in the accuracy of the test, or variability in the patient population. Uncertainty in diagnostic accuracy is common due to poor sensitivity or specificity of a CPR. Significant variability in test results may suggest that the tool is not appropriate for use.

Abbreviations: N, no; NA, not applicable; P, partly; Y, yes.

[a]Items: (1) Prospective design; (2) study site described; (3) justification for the number of participants; (4) representative sample; (5) important patient characteristics described; (6) candidate predictor variables justified; (7) blinded predictor assessment; (8) predictor variables have demonstrated reliability; (9) reference standard valid and reliable; (10) blinded assessment of reference standard; (11) mathematical techniques described; (12) reporting and handling of missing data; (13) 10 outcome events per variable in the final model; (14) 10 outcome events per candidate variable; (15) collinearity of predictor variables assessed; (16) predictor variables kept continuous; (17) uncertainty in CPR accuracy described; (18) uncertainty in post-test probability described; (19) nonparadoxical performance.

TABLE 6 Validation studies—methodological appraisal.

Study (first author)	Items[a]									Total (Y)
	1	2	3	4	5	6	7	8	9	
Benditz et al. (2021)	N	P	N	Y	N	N	N	N	NA	1
Kato et al. (2009)	Y	Y	P	Y	N	Y	Y	Y	NA	6
Kato et al. (2015)	Y	Y	Y	Y	Y	Y	N	Y	NA	7
Tominaga et al. (2022)	Y	Y	Y	Y	Y	N	Y	Y	Y	8

Abbreviations: N, no; NA, not applicable; P, partly; Y, yes.

[a]Items: (1) Prospective validation in new sample; (2) different clinical setting; (3) representative sample; (4) the rule is applied accurately; (5) reliability of the rule is assessed; (6) complete follow-up; (7) reporting and handling of missing data; (8) accuracy uncertainty described; (9) post-test probability uncertainty described.

(31%), primarily due to concerns regarding the reliability and validity of unblinded expert opinions.

3.5 | Results of individual studies

Table 8 summarises the predictive performance of the tools in accordance with the GRASP tool.

The completed GRASP form for each individual study is available in Appendix A3. Out of the 16 identified studies, nine reported the discriminative abilities of the diagnostic CPR to detect the condition of interest using an AUC (ranging from 0.71 to 0.98). Two studies reported the associated AUC confidence interval. The sensitivity and specificity of the diagnostic CPR for ruling in and ruling out a diagnosis was reported in all studies except for Benditz et al. (2019) and Saengdao et al. (2021). The StEP tool (Scholz et al., 2009) demonstrated the best discrimination (AUC = 0.98, $p < 0.001$; sensitivity = 92.0%, specificity = 97.0%) for detecting a diagnosis between radicular and axial LBP. Six of the 11 studies reported the diagnostic CPR model's calibration using the Hosmer-Lemeshow tests (mean = 0.37). Only one of these (Benditz et al., 2019) was considered poorly calibrated with $p < 0.001$.

Table 9 is adapted from Khalifa et al. (2019) and presents the evidence for the 11 diagnostic CPRs based on their stage of development using the GRASP grading of diagnostic prediction tools. All identified diagnostic CPRs underwent internal validation, but only three had external validation (LSS-SSHQ, LSS-DST and StEP), and two of which were externally validated multiple times (LSS-SSHQ and LSS-DST). These diagnostic CPRs, therefore, had the most supporting evidence and were assigned a GRASP grade C1. It should be noted that despite the external validation sample

TABLE 7 Study risk of bias (RoB) and applicability concerns using Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.

	Risk of bias				Applicability concerns		
Study (first author)	Patient selection	Index tests	Reference standard	Flow and timing	Patient selection	Index tests	Reference standard
Konno, Kikuchi, et al. (2007) LSS-SSHQ	Low	Unclear	Unclear	Low	Low	Unclear	Low
LSS-DST Konno, Hayashino, et al. (2007)	Low	Unclear	High	High	Low	Unclear	High
Kato et al. (2009)	Low	Low	High	High	Low	Low	Low
Scholz et al. (2009)	Low	High	Unclear	Low	Low	High	Unclear
Roux et al. (2007)	Low	Unclear	Low	Low	Low	Unclear	Low
Braun et al. (2011)	Low	Low	High	Low	Low	Low	High
Roman et al. (2010)	High	Unclear	Low	Low	High	Unclear	Low
Aghaei et al. (2015)	Low	Low	Unclear	Low	Low	Low	Unclear
Kato et al. (2015) LSS-SSHQ	Low	Unclear	Low	Low	Low	Unclear	Low
Stynes et al. (2018)	High	High	Low	Low	High	High	Low
Benditz et al. (2019)	Low	Low	Low	Low	Low	Low	Low
Saengdao et al. (2021)	High	Unclear	High	Low	High	Unclear	Low
Benditz et al. (2021)	Low	Low	Low	Low	Low	Low	Low
Chatprem et al. (2021)	High	Unclear	Low	Low	Low	Unclear	Low
Ye et al. (2022)	Low	Low	Low	Low	Low	Low	Low
Tominaga et al. (2022)	Low	Low	High	Low	Low	Low	Low

Note: Patient selection: this investigates RoB in aspects of study design like inclusion exclusion criteria, sampling methods, or the definitions used around LBP conditions. Index testing: this explores RoB around the tool itself (i.e. blinding to the result of the reference standard). Reference standard: this explores the validity of the reference standard of diagnosis (i.e. its ability to classify the target condition correctly, or blinding to the index test result). Flow and timing: determines whether all patients received the same reference standard, and if they were included in analysis. It also looks at the time between the reference standard and the index test.

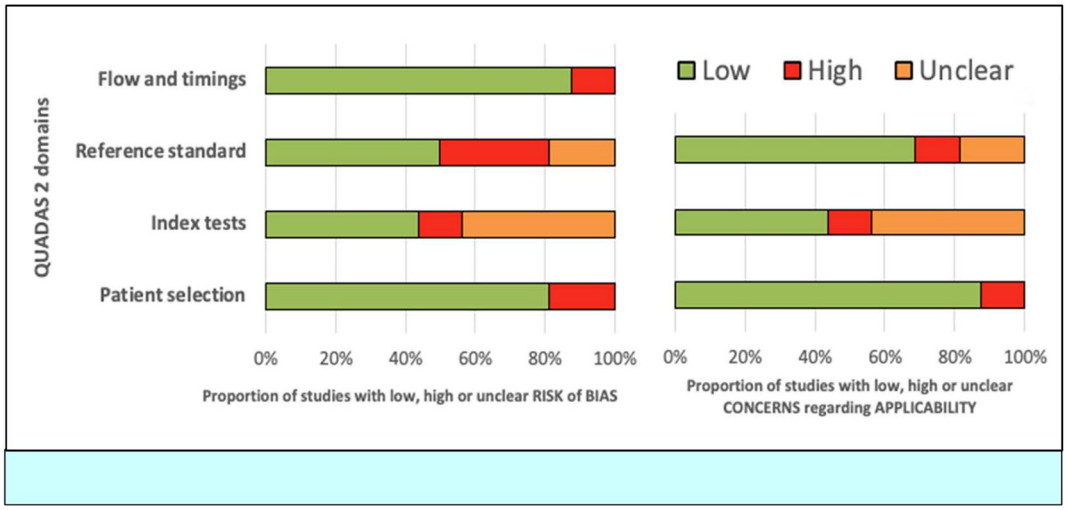

FIGURE 2 Summary of Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) results.

for the LSS-SSHQ being large (n = 33,545), it was still able to demonstrate moderate sensitivity (79.8%) and specificity (68.8%) for detecting lumbar spinal stenosis in primary care (Kato et al., 2015). The LSS-DST was also externally validated multiple times by Kato et al. (2009) and later by Tominaga et al. (2022). Kato reported low specificity (40%); however, the accuracy of this

TABLE 8 Predictive performance of included diagnostic clinical prediction rules (CPRs).

Tool and study (first author)	Discrimination		Calibration
	AUC[a]	Sensitivity %, specificity %	Hosmer–Lemeshow goodness-of-fit
LSS-SSHQ			
Konno, Kikuchi, et al. (2007)	0.80 (derivation)	84.0, 78.0	Not reported
Aghaei et al. (2015)	0.78 (validation)	Radicular type: 97.8, 66.6 Neurogenic type: 97.0, 80.0	Not reported
Kato et al. (2015)	Not reported	79.8, 68.8	Not reported
Tominaga et al. (2022)	Not reported	83.0, 57.0	Not reported
Vert frac			
Roux et al. (2007)	0.77	70.9, 68.6	Not reported
LSS-DST			
Konno, Hayashino, et al. (2007)	0.92	92.8, 72.0	$\chi^{2c} = 11.3$, $p^{b} = 0.19$
Kato et al. (2009)	Not reported	95.0, 40.0	Not reported
Tominaga et al. (2022)	Not reported	91.0, 76.0	Not reported
StEP			
Scholz et al. (2009)	0.98	92.0, 97.0	Not reported
OVCF			
Roman et al. (2010)	Not reported	37.0, 96.0	Not reported
AxSpA			
Braun et al. (2011)	0.71	47.8, 86.1	Not reported
Keele SCIATICA			
Stynes et al. (2018)	0.95	85.0, 88.0	$\chi^2 = 11.4$, $p = 0.18$
Regensburg			
Benditz et al. (2019)	Not reported	Not reported	$\chi^2 = 10.5$, $p < 0.001$
Benditz et al. (2021)	Not reported	Not reported	Not reported
Lx inst			
Chatprem et al. (2021)	0.78	5.6, 99.0	$p = 0.33$
Occ-LBP			
Saengdao et al. (2021)	0.90	Not reported	$p = 0.67$
Clinical nomogram			
Ye et al. (2022)	0.90	93.9, 62.4	$p = 0.85$

[a](Area under ROC curve).
[b]Probability.
[c]Chai squared.

result was limited by a small sample size ($n = 118$) and high RoB due to poor agreement around the reference standard for LSS. The specificity of the LSS-DST was later reported as 76% in a larger ($n = 3331$), more robust study from Tominaga et al. (2022). The study reported excellent discrimination for detecting axial and radicular LBP (AUC 0.98; $p < 0.001$) and showed high usability between patients and clinicians. This was assigned a GRASP grade B1. Tools developed before 2018 were cited more frequently (mean citations, $n = 109$) compared to newer tools (mean citations, $n = 2$). Despite not being widely reported in other studies, the diagnostic CPRs developed by Stynes et al. (2018), Saengdao et al. (2021), and Ye et al. (2022) have all shown strong positive evidence and low risk of bias during the early stages of development.

TABLE 9 Summary of Grade and Assess Predictive tool (GRASP) grading of 11 diagnostic clinical prediction rules (CPRs).

Tools	Tool Information					Impact After Implementation			During Implementation		Predictive Performance Before Implementation		
	Country	Year	Citations	Studies	Tool Grade assigned (CH)	Experimental Studies	Observational Studies	Subjective Studies	Usability	Potential Effect	External Validation Multiple Times	External Validation Only Once	Internal Validation
						A1	A2	A3	B1	B2	C1	C2	C3
LSS-SSHQ	Japan	2007	108	3	**C1**						●	●	●
Vert Frac	France	2007	47	1	C3								◔
LSS-DST	Japan	2007	107	3	C1						●	◔	◔
StEP	USA & UK	2009	311	2	B1				●			●	●
OCVF	USA	2010	24	1	C3								●
AxSpA	Germany	2011	131	1	C3								●
Keele SCIATICA	UK	2018	34	1	C3								●
Regensburg	Germany	2019	4	2	C3								◔
Lx Inst	Thailand	2021	2	1	C3								◔
Occ-LBP	Japan	2021	0	1	C3								●
Clinical Nomogram	China	2022	2	1	C3								●
Evidence Direction	Positive Evidence ●				Mixed Evidence Supporting Positive Conclusion ◔								
	Negative Evidence ○				Mixed Evidence Supporting Negative Conclusion ◔								

4 | DISCUSSION

4.1 | Summary of findings

This review aimed to summarise LBP-related diagnostic CPRs and evaluate their readiness for implementation in clinical practice. To the authors' knowledge, this is the only systematic review since Haskins et al. (2015) to have synthesised the evidence of LBP diagnostic CPRs. In total, 11 diagnostic CPRs were identified in this review, six of which were previously identified by Haskin et al. (2012) and again by Haskins et al. (2015), and five were more recently developed. In keeping with the conclusions made by Haskins et al. (2015), this review also found that the majority of these diagnostic CPRs have not yet been developed past the initial derivation phase (n = 8/11), and therefore cannot be recommended for implementation into clinical practice at present. Three identified diagnostic CPRs were externally validated at least once and have shown moderate utility in providing diagnostic support in patients with LSS (LSS-SSHQ, LSS-DST) or axial/radicular type LBP (StEP). The StEP tool was the only diagnostic CPR to report its usability, being deemed 'highly useable' by patients (n = 134) and having high clinical face-validity. There were no differences in the usability of the StEP tool when used with either axial or radicular LBP patients (Scholz et al., 2009). The clinical impact of these three diagnostic CPRs remains unclear.

4.2 | Results in the context of other evidence

The lack of validation studies highlighted in this review is not uncommon for clinical prediction rules (Steyerberg, 2009). Several papers have highlighted a concerning trend of academic teams developing new CPRs instead of externally validating or updating existing ones (Binuya et al., 2022; Bouwmeester et al., 2012; Hendriksen et al., 2013; Moons et al., 2009), with even fewer studies properly testing their impact on clinical decision-making or patient outcomes (Reily et al., 2003). It is therefore not surprising that the overwhelming majority of developed prediction models are not used in practice when there is a lack of external validation studies describing their performance in external samples (van Calster et al., 2023).

Inadequate reporting of sample size justification and procedures for managing missing data is a persistent issue in this field, despite

the emphasis on these factors in various reporting guidelines (Bouwmeester et al., 2012). The present study, similarly identified that sample size justification (3/11) and missing data processes (2/11) were also poorly reported, highlighting ongoing concerns with these methodological issues. In a large systematic review of over 120 non-LBP predictive models (Collins et al., 2014), the calibration (the agreement between the predicted probabilities from the CPR and the outcome of interest, i.e., LSS) was deemed poorly reported in studies, despite being a key measure of CPR performance. This held true in the current study for the earlier diagnostic CPRs identified (2007–2011), in which only 1/6 studies reported calibration. However, between 2011 and 2022, a significant improvement in calibration reporting was seen, with all five studies reporting calibration adequately. This may be attributed to the publication of the TRIPOD (transparent reporting of multivariable prediction models for individual prognosis or diagnosis) statement developed by Collins et al. (2015), which includes 'calibration' as one of 22 essential items on their checklist designed to guide the reporting of prediction model studies.

It should be noted that prognostic LBP tools, such as STarT Back (Hill et al., 2008) and Orebro (Linton & Halldén, 1998), have not only been extensively externally validated in different settings but also undergone experimental and observational study designs (GRASP grade A1). By comparison, these diagnostic CPRs have so far only undergone usability testing (GRASP grade B1).

One finding from this RoB appraisal (QUADAS-2) was the high RoB concerning the reference standard, especially in tools related to the diagnosis of LSS ($n = 3/5$ tools with high reference standard RoB). Due to poor reporting of blinding and robustness of the reference standard used, the clinical utility of the LSS-DST and LSS-SSHQ remains uncertain. For example, some studies used the attending physician's impression, whereas others were more thorough and used a panel of orthopaedic surgeons to confirm diagnosis. In the absence of a gold standard for diagnosing LSS (Cook et al., 2020), clinicians should exercise caution when interpreting the diagnostic performance of these tools.

4.3 | Limitations

This review used an operational definition of a diagnostic CPR designed to reflect the most common use of the term 'CPR' in the literature, considering definitions from other relevant systematic reviews (Haskins et al., 2012, 2015). Eligible studies were sensitive to meeting this operational definition, which was the largest reason for exclusion after full text screening. For example, this definition excluded studies deemed to be 'clinical classification criteria' or 'screening tools', because the American College of Rheumatology highlights that the purpose of these tools is not to support diagnostic decision-making (Aggarwal et al., 2015). Due to these recommendations, the 'ASAS classification criteria' for ankylosing spondylitis (Sieper et al., 2009), and the 'chronic LBP screening tool' for patients who might benefit from psychological assessment (Apeldoorn et al., 2012), were both excluded despite being previously labelled as diagnostic CPRs by Haskins et al. (2015). Alternative operational definitions of diagnostic CPRs would likely have led to different tools being included in the evidence synthesis.

It should be noted that in the absence of a validated methodological quality appraisal tool for CPRs (Binuya et al., 2022), the methodological quality appraisal items used in this study were based on well-cited CPR methodological standards (Cowley et al., 2019), and recent CPR systematic reviews. Furthermore, the methodological quality (GRASP) and RoB appraisals were performed by a single reviewer (CH), increasing the risk of reviewer or confirmational biases (Drucker et al., 2016); therefore, findings should be interpreted with caution.

4.4 | Implications

This evidence synthesis found five additional diagnostic CPRs that have been developed since the Haskins et al. (2015) review; however, none of these have been sufficiently validated for use in clinical practice. Three diagnostic CPRs (LSS-SSHQ, LSS-DST and StEP) previously identified by Haskins et al. (2015) have now been externally validated, with eight still requiring further validation. In the absence of impact analysis studies, caution remains regarding the use of the LSS-SSHQ, LSS-DST and StEP in clinical practice. Additional external validation and impact analysis studies are warranted for existing tools with robust supporting evidence. This review also highlights the need for future research to abide by reporting methodological guidelines (such as the TRIPOD statement) for derivation and validation studies as the methodological quality in some studies was limited.

4.5 | Conclusions

This is the first systematic review since Haskins et al. (2015) to have synthesised the evidence of LBP diagnostic CPRs. Five new diagnostic CPRs have been developed, but all remain in the early phase of testing. Since 2015, two diagnostic CPRs for lumbar spinal stenosis (LSS-SSHQ and LSS-DST) and one diagnostic CPR to distinguish between radicular and axial only LBP (StEP) have been externally validated. Further research is needed to determine the impact of these tools on clinical decision-making and patient outcomes, and in the meantime, their clinical use should be considered with caution. Eight further diagnostic CPRs show initial promise for supporting decision-making around other LBP diagnoses but are not yet ready for clinical implementation.

AUTHOR CONTRIBUTIONS

Charles James Hill (primary author): Designed the research question and study aims, wrote the original manuscript, screened the studies and synthesised the data. Claire Stapleton (supervisor and marker): Provided invaluable supervision and guidance throughout the review

process. Claire Stapleton also served as the primary supervisor and marker for the project, offering critical insights and oversight. Anirban Banerjee (secondary marker): Contributed as a secondary marker, providing additional evaluation and feedback to ensure the quality and rigour of the systematic review. Jonathan Hill (manuscript refinement): Made significant contributions to the editing and refinement of the manuscript, enhancing its clarity and coherence.

ACKNOWLEDGEMENTS

I extend my heartfelt appreciation to Patty Guan for her meticulous article screening, which greatly enriched the quality of this systematic review. I am also grateful to my supervisor, Dr Claire Stapleton, for her unwavering guidance and support throughout this project (Appendix A5). Both of their contributions were invaluable in shaping the final outcome of this work.

CONFLICT OF INTEREST STATEMENT

All authors declare no conflict-of-interest present.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ETHICS STATEMENT

Transparency and Integrity: We are committed to conducting this systematic review using open and honest research processes to uphold transparency and integrity throughout. This extends to the accurate and comprehensive reporting of all data, methods and findings. **Objectivity:** This review will be approached objectively, and impartially, taking care to avoid any bias in the selection, analysis, and presentation of the included studies. Any potential conflicts of interest among our reviewers will be fully disclosed. **Inclusion and Exclusion Criteria:** We will clearly define and document our inclusion and exclusion criteria. The selection of studies for inclusion will be based solely on their relevance to the research question, without consideration of external factors. **Plagiarism and Attribution:** We will adhere to strict standards of academic integrity, providing proper attribution to the authors of studies included in the review and avoiding plagiarism throughout. **Publication Bias:** To mitigate publication bias, we will conduct comprehensive searches of major literature databases, following the same standards as other relevant published systematic reviews. Any potential bias in our findings will be transparently reported in the review. **Peer review:** We actively encourage the peer review process to optimise the quality and rigour of our research. We commit to address any feedback received from peer reviewers in a timely and appropriate manner. **Continuous Improvement:** We welcome feedback from the scientific community and stakeholders. Such input will be carefully considered for potential updates or revisions of this systematic review. By adhering to these ethical principles and guidelines, we aim to conduct a systematic review that upholds the highest standards of research integrity and contributes to the advancement of knowledge in the field of low back pain diagnosis.

ORCID

Charles James Hill https://orcid.org/0009-0007-1284-9123

REFERENCES

Aggarwal, R., Ringold, S., Khanna, D., Neogi, T., Johnson, S. R., Miller, A., Brunner, H. I., Ogawa, R., Felson, D., Ogdie, A., Aletaha, D., & Feldman, B. M. (2015). Distinctions between diagnostic and classification criteria? *Arthritis Care & Research*, *67*(7), 891–897. https://doi.org/10.1002/acr.22583

Aghaei, N. H., Azimi, P., Shahzadi, S., Azhari, S., & Mohammadi, H. R. (2015). Role of the self-administered, self-reported history questionnaire to identify types of lumbar spinal stenosis: A sensitivity analysis. *Asian Spine Journal*, *9*(5), 689–693. https://doi.org/10.4184/asj.2015.9.5.689

Apeldoorn, A. T., Bosselaar, H., Ostelo, R. W. J. G., Blom-Luberti, T., van der Ploeg, T., Fritz, J. M., de Vet, H. C. W., & van Tulder, M. W. (2012). Identification of patients with chronic low back pain who might benefit from additional psychological assessment. *The Clinical Journal of Pain*, *28*(1), 23–31. https://doi.org/10.1097/ajp.0b013e31822019d0

Beattie, P., & Nelson, R. (2006). Clinical prediction rules: What are they and what do they tell us? *Australian Journal of Physiotherapy*, *52*(3), 157–163. https://doi.org/10.1016/s0004-9514(06)70024-1

Benditz, A., Faber, F., Wenk, G., Fuchs, T., Salak, N., Grifka, J., Vogl, M., Menke, M., & Jansen, P. (2019). The role of a decision support system in back pain diagnoses: A pilot study. *BioMed Research International*, *13140*(2), 8–15.

Benditz, A., Pulido, L. C., Grifka, J., Ripke, F., & Jansen, P. (2021). A clinical decision support system in back pain helps to find the diagnosis: A prospective correlation study. *Archives of Orthopaedic and Trauma Surgery*, *143*(2), 621–625. https://doi.org/10.1007/s00402-021-04080-y

Beneciuk, J. M., Bishop, M. D., & George, S. Z. (2009). Clinical prediction rules for physical therapy interventions: A systematic review. *Physical Therapy*, *89*(2), 114–124. https://doi.org/10.2522/ptj.20080239

Binuya, M. A. E., Engelhardt, E. G., Schats, W., Schmidt, M. K., & Steyerberg, E. W. (2022). Methodological guidance for the evaluation and updating of clinical prediction models: A systematic review. *BMC Medical Research Methodology*, *22*(1), 316. https://doi.org/10.1186/s12874-022-01801-8

Bouwmeester, W., Zuithoff, N., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E., Altman, D. G., & Moons, K. (2012). Reporting and methods in clinical prediction research: A systematic review. *PLoS Medicine*, *9*(5), 1–12. https://doi.org/10.1371/journal.pmed.1001221

Braun, A., Saracbasi, E., Grifka, J., Schnitker, J., & Braun, J. (2011). Identifying patients with axial spondyloarthritis in primary care: How useful are items indicative of inflammatory back pain? *Annals of the Rheumatic Diseases*, *70*(10), 1782–1787. https://doi.org/10.1136/ard.2011.151167

Chatprem, T., Puntumetakul, R., Kanpittaya, J., Selfe, J., & Yeowell, G. (2021). A diagnostic tool for people with lumbar instability: A criterion-related validity study. *BMC Musculoskeletal Disorders*, *22*(1), 1–976. https://doi.org/10.1186/s12891-021-04854-w

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L., Moons, K. G., & Altman, D. G. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*(1), 40. https://doi.org/10.1186/1471-2288-14-40

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, *162*(1), 55–63. https://doi.org/10.7326/m14-0697

Cook, C. (2008). Potential pitfalls of clinical prediction rules. *Journal of Manual & Manipulative Therapy*, *16*(2), 69–71. https://doi.org/10.1179/106698108790818477

Cook, C. J., Cook, C. E., Reiman, M. P., Joshi, A. B., Richardson, W., & Garcia, A. N. (2020). Systematic review of diagnostic accuracy of patient history, clinical findings, and physical tests in the diagnosis of lumbar spinal stenosis. *European Spine Journal*, *29*(1), 93–112. https://doi.org/10.1007/s00586-019-06048-4

Cowley, L. E., Farewell, D. M., Maguire, S., & Kemp, A. M. (2019). Methodological standards for the development and evaluation of clinical prediction rules: A review of the literature. *Diagnostic and prognostic research*, *3*, 16. https://doi.org/10.1186/s41512-019-0060-y

Drucker, A. M., Fleming, P., & Chan, A. (2016). Research techniques made simple: Assessing risk of bias in systematic reviews. *Journal of Investigative Dermatology*, *136*(11), e109–e114. https://doi.org/10.1016/j.jid.2016.08.021

Foster, N. E., Anema, J. R., Cherkin, D., Chou, R., Cohen, S. P., Gross, D. P., Ferreira, P. H., Fritz, J. M., Koes, B. W., Peul, W., Turner, J. A., Maher, C. G., Buchbinder, R., Hartvigsen, J., Foster, N. E., Underwood, M., van Tulder, M., & Woolf, A. (2018). Prevention and treatment of low back pain: Evidence, challenges, and promising directions. *The Lancet (British Edition)*, *391*(10137), 68–83. https://doi.org/10.1016/s0140-6736(18)30489-6

Geersing, G., Bouwmeester, W., Zuithoff, P., Spijker, R., Leeflang, M., Moons, K. G., & Moons, K. (2012). Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One*, *7*(2), e32844. https://doi.org/10.1371/journal.pone.0032844

Hartvigsen, J., Hancock, M. J., Kongsted, A., Louw, Q., Ferreira, M. L., Genevay, S., Hoy, D., Karppinen, J., Pransky, G., Sieper, J., Smeets, R. J., Underwood, M., Buchbinder, R., Cherkin, D., Foster, N. E., Maher, C. G., Underwood, M., van Tulder, M., & Woolf, A. (2018). What low back pain is and why we need to pay attention. *The Lancet (British Edition)*, *391*(10137), 2356–2367. https://doi.org/10.1016/s0140-6736(18)30480-x

Haskins, R., Osmotherly, P. G., & Rivett, D. A. (2015). Diagnostic clinical prediction rules for specific subtypes of low back pain: A systematic review. *Journal of Orthopaedic & Sports Physical Therapy*, *45*(2), 61–76. https://doi.org/10.2519/jospt.2015.5723

Haskins, R., Rivett, D. A., & Osmotherly, P. G. (2012). Clinical prediction rules in the physiotherapy management of low back pain: A systematic review. *Manual Therapy*, *17*(1), 9–21. https://doi.org/10.1016/j.math.2011.05.001

Hebert, J. J., Koppenhaver, S. L., & Walker, B. F. (2011). Subgrouping patients with low back pain. *Sport Health*, *3*(6), 534–542. https://doi.org/10.1177/1941738111415044

Heniksen, J. T., Geersing, G. J., Moons, K. G. M., & de Groot, J. A. (2013). Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis*, *11*(1), 129–141. https://doi.org/10.1111/jth.12262

Hill, J. C., Dunn, K. M., Lewis, M., Mullis, R., Main, C. J., Foster, N. E., & Hay, E. M. (2008). A primary care back pain screening tool: Identifying patient subgroups for initial treatment. *Arthritis & Rheumatism*, *59*(5), 632–641. https://doi.org/10.1002/art.23563

Institute for Health Metrics and Evaluation. (2020). *GBD compare data visualization*. IHME, University of Washington. [Google Scholar]. Retrieved from: http://vizhub.healthdata.org/gbd-compare. Accessed 2 July 2023.

Jensen, R. K., Lauridsen, H. H., Andresen, A. D. K., Mieritz, R. M., Schiøttz-Christensen, B., & Vach, W. (2020). Diagnostic screening for lumbar spinal stenosis. *Clinical Epidemiology*, *12*, 891–905. https://doi.org/10.2147/clep.s263646

Kato, K., Sekiguchi, M., Yonemoto, K., Kakuma, T., Nikaido, T., Watanabe, K., Otani, K., Yabuki, S., Kikuchi, S., & Konno, S. (2015). Diagnostic accuracy of the self-administered, self-reported history questionnaire for lumbar spinal stenosis patients in Japanese primary care settings: A multicenter cross-sectional study (DISTO-project). *Journal of Orthopaedic Science*, *20*(5), 805–810. https://doi.org/10.1007/s00776-015-0740-6

Kato, Y., Kawakami, T., Kifune, M., Kishimoto, T., Nibu, K., Oda, H., Shirasawa, K., Tominaga, T., Toyoda, K., Tsue, K., & Taguchi, T. (2009). Validation study of a clinical diagnosis support tool for lumbar spinal stenosis. *Journal of Orthopaedic Science*, *14*(6), 711–718. https://doi.org/10.1007/s00776-009-1391-2

Khalifa, M., Magrabi, F., & Gallego, B. (2019). Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support. *BMC Medical Informatics and Decision Making*, *19*(1), 207. https://doi.org/10.1186/s12911-019-0940-7

Konno, S., Hayashino, Y., Fukuhara, S., Kikuchi, S., Kaneda, K., Seichi, A., Chiba, K., Satomi, K., Nagata, K., & Kawai, S. (2007). Development of a clinical diagnosis support tool to identify patients with lumbar spinal stenosis. *European Spine Journal*, *16*(11), 1951–1957. https://doi.org/10.1007/s00586-007-0402-2

Konno, S., Kikuchi, S., Tanaka, Y., Yamazaki, K., Shimada, Y., Takei, H., Yokoyama, T., Okada, M., & Kokubun, S. (2007). A diagnostic support tool for lumbar spinal stenosis: A self-administered, self-reported history questionnaire. *BMC Musculoskeletal Disorders*, *8*(1), 102. https://doi.org/10.1186/1471-2474-8-102

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Linton, S., & Halldén, K. (1998). Can we screen for problematic back pain? A screening questionnaire for predicting outcome in acute and subacute back pain. *The Clinical Journal of Pain*, *14*(3), 209–215. https://doi.org/10.1097/00002508-199809000-00007

Lubetzky-Vilnai, A., Ciol, M., & McCoy, S. (2014). Statistical analysis of clinical prediction rules for rehabilitation interventions: Current state of the literature. *Archives of Physical Medicine and Rehabilitation*, *95*(1), 188–196. https://doi.org/10.1016/j.apmr.2013.08.242

May, S., & Rosedale, R. (2009). Prescriptive clinical prediction rules in back pain research: A systematic review. *Journal of Manual & Manipulative Therapy*, *17*(1), 36–45. https://doi.org/10.1179/106698109790818214

McGinn, T., Jervis, R., Wisnivesky, J., Keitz, S., & Wyer, P. C. (2008). Tips for teachers of evidence-based medicine: Clinical prediction rules (CPRs) and estimating pretest probability. *Journal of General Internal Medicine*, *23*(8), 1261–1268. https://doi.org/10.1007/s11606-008-0623-z

McGinn, T. G., Guyatt, G. H., Wyer, P. C., Naylor, C. D., Stiell, I. G., Richardson, W. S., & for the Evidence-Based Medicine Working Group (2000). Users' guides to the medical literature: XXII: How to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA, the Journal of the American Medical Association*, *284*(1), 79–84. https://doi.org/10.1001/jama.284.1.79

Microsoft Corporation. (2023). Microsoft Excel. Retrieved from https://office.microsoft.com/excel

Moons, K. G. M., de Groot, J. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. *PLoS Medicine*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

Moons, K. M., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: What, why, and how? *BMJ*, *338*(7706), 1317–1320. https://doi.org/10.1136/bmj.b375

Munn, Z., Peters, M. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

NHS England. (2014). Trauma programme of care pathfinder project – low back pain and radicular pain. Retrieved from: https://rcc-uk.org/wp-

content/uploads/2015/03/Pathfinder-Low-back-and-Radicular-Pain_Final.pdf

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., …, & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (Online)*, *372*, n71. https://doi.org/10.1136/bmj.n71

Patel, S., Friede, T., Froud, R., Evans, D. W., & Underwood, M. (2013). Systematic review of randomized controlled trials of clinical prediction rules for physical therapy in low back pain. *Spine*, *38*(9), 762–769. https://doi.org/10.1097/brs.0b013e31827b158f

Petersen, T., Laslett, M., & Juhl, C. (2017). Clinical classification in low back pain: Best-evidence diagnostic rules based on systematic reviews. *BMC Musculoskeletal Disorders*, *18*(1), 188. https://doi.org/10.1186/s12891-017-1549-6

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., Jones, D. R., Henry, D., & Burchill, S. A. (2003). Reporting of prognostic markers: Current problems and development of guidelines for evidence-based practice in the future. *British Journal of Cancer*, *88*(8), 1191–1198. https://doi.org/10.1038/sj.bjc.6600886

Roman, M., Brown, C., Richardson, W., Isaacs, R., Howes, C., & Cook, C. (2010). The development of a clinical decision-making algorithm for detection of osteoporotic vertebral compression fracture or wedge deformity. *Journal of Manual & Manipulative Therapy*, *18*(1), 44–49. https://doi.org/10.1179/106698110x12595770849641

Roux, C., Priol, G., Fechtenbaum, J., Cortet, B., Liu-Léage, S., & Audran, M. (2007). A clinical tool to determine the necessity of spine radiography in postmenopausal women with osteoporosis presenting with back pain. *Annals of the Rheumatic Diseases*, *66*(1), 81–85. https://doi.org/10.1136/ard.2006.051474

Saengdao, O., Surasak, B., & Jayanton, P. (2021). A diagnostic assistant tool for work-related low back pain in hospital workers. *Indian Journal of Occupational and Environmental Medicine*, *25*(1), 11–16. https://doi.org/10.4103/ijoem.ijoem_153_19

Saragiotto, B. T., Maher, C. G., Hancock, M. J., & Koes, B. (2017). Subgrouping patients with nonspecific low back pain: Hope or hype? *Journal of Orthopaedic & Sports Physical Therapy*, *47*(2), 44–48. https://doi.org/10.2519/jospt.2017.0602

Scholz, J., Mannion, R. J., Hord, D. E., Griffin, R. S., Rawal, B., Zheng, H., Scoffings, D., Phillips, A., Guo, J., Laing, R. J. C., Abdi, S., Decosterd, I., & Woolf, C. J. (2009). A novel tool for the assessment of pain: Validation in low back pain. *PLoS Medicine*, *6*(4), e1000047. https://doi.org/10.1371/journal.pmed.1000047

Scott, I. A., & Crock, C. (2020). Diagnostic error: Incidence, impacts, causes, and preventive strategies. *Medical Journal of Australia*, *213*(7), 302–305. https://doi.org/10.5694/mja2.50771

Seel, R. T., Steyerberg, E. W., Malec, J. F., Sherer, M., & Macciocchi, S. N. (2012). Developing and evaluating prediction models in rehabilitation populations. *Archives of Physical Medicine and Rehabilitation*, *93*(8), S138–S153. https://doi.org/10.1016/j.apmr.2012.04.021

Sieper, J., van der Heijde, D., Landewé, R., Brandt, J., Burgos-Vagas, R., Collantes-Estevez, E., Dijkmans, B., Dougados, M., Khan, M. A., Leirisalo-Repo, M., van der Linden, S., Maksymowych, W. P., Mielants, H., Olivieri, I., & Rudwaleit, M. (2009). New criteria for inflammatory back pain in patients with chronic back pain: A real patient exercise by experts from the Assessment of SpondyloArthritis international Society (ASAS). *Annals of the Rheumatic Diseases*, *68*(6), 784–788. https://doi.org/10.1136/ard.2008.101501

Stanton, T. R., Hancock, M. J., Maher, C. G., & Koes, B. (2010). Critical appraisal of clinical prediction rules that aim to optimize treatment selection for musculoskeletal conditions. *Physical Therapy*, *90*(6), 843–854. https://doi.org/10.2522/ptj.20090233

Steyerberg, E. W. (2009). *Clinical prediction models* (1st ed.). Springer-Verlag.

Stynes, S., Konstantinou, K., Ogollah, R., Hay, E. M., & Dunn, K. M. (2018). Clinical diagnostic model for sciatica developed in primary care patients with low back-related leg pain. *PLoS One*, *13*(4), e0191852. https://doi.org/10.1371/journal.pone.0191852

Tominaga, R., Kurita, N., Sekiguchi, M., Yonemoto, K., Kakuma, T., & Konno, S. (2022). Diagnostic accuracy of the lumbar spinal stenosis-diagnosis support tool and the lumbar spinal stenosis-self-administered, self-reported history questionnaire. *PLoS One*, *17*(5), e0267892. https://doi.org/10.1371/journal.pone.0267892

Van Calster, B., Steyerberg, E. W., Wynants, L., & van Smeden, M. (2023). There is no such thing as a validated prediction model. *BMC Medicine*, *21*(1), 70. https://doi.org/10.1186/s12916-023-02779-w

van Geloven, N., Giardiello, D., Bonneville, E. F., Teece, L., Ramspek, C. L., van Smeden, M., Snell, K. I. E., van Calster, B., Pohar-Perme, M., Riley, R. D., Putter, H., & Steyerberg, E. (2022). Validation of prediction models in the presence of competing risks: A guide through modern methods. *BMJ (Online)*, *377*, e069249. https://doi.org/10.1136/bmj-2021-069249

van Oort, L., van den Berg, T., Koes, B. W., de Vet, R. H. C. W., Anema, H. J. R., Heymans, M. W., & Verhagen, A. P. (2012). Preliminary state of development of prediction models for primary care physical therapy: A systematic review. *Journal of Clinical Epidemiology*, *65*(12), 1257–1266. https://doi.org/10.1016/j.jclinepi.2012.05.007

Whiting, P. F., Rutjes, A. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Ye, L., Miao, S., Xiao, Q., Liu, Y., Tang, H., Li, B., Liu, J., & Chen, D. (2022). A predictive clinical-radiomics nomogram for diagnosing axial spondyloarthritis using MRI and clinical risk factors. *Rheumatology*, *61*(4), 1440–1447. https://doi.org/10.1093/rheumatology/keab542

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hill, C. J., Banerjee, A., Hill, J., & Stapleton, C. (2023). Diagnostic clinical prediction rules for categorising low back pain: A systematic review. *Musculoskeletal Care*, 1–15. https://doi.org/10.1002/msc.1816