**Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews.**

\*Scott D Tagliaferri 1 MSc ClinExPhys

Maia Angelova 2 *PhD*

Xiaohui Zhao 3 *PhD*

Patrick J Owen 1 *PhD*

Clint T Miller 1 *PhD*

Tim Wilkin 2 *PhD*

\*Daniel L Belavy 1 *PhD*

1 Deakin University, Institute for Physical Activity and Nutrition (IPAN), School of Exercise and Nutrition Sciences, Geelong, Victoria, Australia

2 Deakin University, School of Information Technology, Geelong, Victoria, Australia

3 Xi’an University of Architecture & Technology, Beilin, Xi'an, China

**Corresponding author (whilst in review):** Scott Tagliaferri; Deakin University, Institute for Physical Activity and Nutrition, School of Exercise and Nutrition Sciences, 221 Burwood Highway, Burwood, Victoria, 3125, Australia; Tel: +61 3 9244 6606; Email: scott.tagliaferri@deakin.edu.au

**Corresponding author (upon acceptance)**: A/Prof Daniel L Belavy, PhD; Deakin University, Institute for Physical Activity and Nutrition, School of Exercise and Nutrition Sciences, 221 Burwood Highway, Burwood, Victoria, 3125, Australia. Tel: +61 3 9244 6606; Fax: +61 3 9244 6017; E-mail: d.belavy@deakin.edu.au, belavy@gmail.com.

**Short title**: Artificial intelligence and clinical classification in back pain

**Word count**: 4635

**ABSTRACT**

Artificial intelligence and machine learning (AI/ML) could enhance the ability to detect patterns of clinical characteristics in low back pain (LBP) and guide treatment. We conducted three systematic reviews to address the following aims: (a) review the status of AI/ML research in LBP, (b) compare its status to that of two established LBP classification systems (STarT Back, McKenzie). AI/ML in LBP is in its infancy: 45 of 48 studies assessed sample sizes <1000 people, 19 of 48 studies used ≤5 parameters in models, 13 of 48 studies applied multiple models and attained high accuracy, 25 of 48 studies assessed the binary classification of LBP versus no-LBP only. Beyond the 48 studies using AI/ML for LBP classification, no studies examined use of AI/ML in prognosis prediction of specific sub-groups, and AI/ML techniques are yet to be implemented in guiding LBP treatment. In contrast, the STarT Back tool has been assessed for internal consistency, test-retest reliability, validity, pain and disability prognosis, and influence on pain and disability treatment outcomes. McKenzie has been assessed for inter- and intra-tester reliability, prognosis, and impact on pain and disability outcomes relative to other treatments. For AI/ML methods to contribute to the refinement of LBP (sub‑)classification and guide treatment allocation, large data sets containing known and exploratory clinical features should be examined. There is also a need to establish reliability, validity, and prognostic capacity of AI/ML techniques in LBP as well as its ability to inform treatment allocation for improved patient outcomes and/or reduced health care costs.

**Key words**: back pain; spine; artificial intelligence; machine learning

**INTRODUCTION**

Low back pain (LBP) is the leading cause of disability worldwide1 and is associated with annual economic costs up to AU $9·2 billion2 and US $102 billion3 in Australia and the United States of America, respectively. In addition to economic burden, multiple individual factors (e.g. loss of social identity,4 distress5 and physical deconditioning6) contribute to pain intensity and disability in this population group.7 Approximately 90% of people with LBP are classified as having ‘non-specific’ LBP, where no clear tissue cause of pain can be found.8 However, we anticipate that people with non-specific LBP are not a homogeneous group, yet the challenge remains to identify potential sub-groups that could benefit from specific treatments to assist in reducing the burden of the condition.9

Artificial intelligence and machine learning (AI/ML) techniques have been used to improve the understanding, diagnosis and management of acute and chronic diseases.10 Technological advancements, such as machine learning algorithms, have led to an increased capacity to recognise patterns in data sets, and used successfully to classify individuals with liver disease and heart failure10,11 and have found some application more widely in pain research.12 However, the utilisation of such techniques in LBP, to date, is limited. The primary aim of this work was to conduct a systematic review examining how machine learning tools have been used in LBP.

A classification approach or assessment tool that is implemented in clinical practice should have utility: be it for the patient (e.g. improved outcomes) and/or for the health care system (e.g. reduced costs). Any classification tool should ideally be (a) reliable, (b) valid, (c) detect people who are likely to have a different outcome or prognosis and (d) its implementation in clinical practice should improve patient outcomes overall and/or reduce health care costs and reduce the burden of disease.13-15 To illustrate the current status, and potential future direction, of AI/ML approaches to LBP, we contrasted this to two commonly implemented clinical classification approaches (McKenzie16 and STarT Back13). The McKenzie method has been extensively studied in randomised clinical trials (RCTs) and subsequent meta-analyses of LBP treatment,17 while the STarT Back tool is currently recommended in national guidelines.18 McKenzie is a classification method of diagnosing movement preferences (e.g. spinal extension versus flexion) based on symptoms response (e.g. centralisation versus peripheralization of symptoms),16 while the STarT Back classifies people in to low, medium and high risk of developing persistent disabling symptoms based on physical and psychosocial factors.13 A comparison of AI/ML utilisation to these existing clinical classification approaches can guide future work in sub-classification of LBP using AI/ML; specifically allowing for the development of a more robust tool that has the potential to impact the burden of disease of LBP. Therefore, (a) the primary aim was to systematically review the literature on AI/ML in LBP research, (b) while a secondary aim was to systematically review and contrast two common LBP classification approaches that are in active use in clinical practice (McKenzie and STarT Back) to how AI/ML tools have been used to date. To do this, we considered the reliability, validity, and prognostic capacity of these classification systems, as well as their impact on patient outcomes (e.g. pain intensity and disability) and healthcare costs, as determined in RCTs.

**RESULTS**

**Machine learning**

Despite broad search terms, only 185 articles were identified after duplicate removal, with 64 assessed at the full-text stage (Figure 1). The reasons for exclusion of AI/ML studies at the full-text stage are presented in Supplementary Table 7. A total of 48 studies were included in data extraction and qualitative synthesis.19-66 Reasons for study exclusion are provided in Figure 1.

The overview of study characteristics and authors conclusions is presented in Table 1. Studies were split into case‑control, cohort or other classifications. Overall, the sample sizes range from 10 to 34,589 people. The populations consisted of 16 studies that looked at chronic LBP,19,20,24,28,29,31,36,37,39,42,54-57,62,64 two acute LBP,27,30 one recurrent,22 one lumbar spinal stenosis,21 two surgical,46,61 nine other (mixed samples)35,38,40,41,48,51,53,65,66 and 17 were unclear (LBP type not defined).23,25,26,32-34,43-45,47,49,50,52,58-60,63 Ten studies did not report training and testing of the data sets.26,29,33,46,51,52,55,56,59,60

Classification of LBP was assessed in 25 studies, all of which attempted binary classification to detect the presence of LBP or not.19,20,23-25,28,29,31-33,37,40-42,44,47,49,50,53-55,57,62-64 One study classified golfers with and without LBP based on electromyography and golf kinematic data using a Support Vector Machine (multilayer perceptron with one layer, where input data are placed into vector spaces)12 with 100% accuracy.47 Another study looked at classifying LBP based on the number of contacts with healthcare professionals with an accuracy of 91%.34 Four studies23,32,40,41 classified LBP and controls based on electromyography, spinal positions and trunk range of motion. Sample sizes of these studies range from 98 to 1,510. The accuracy of these studies for classifying LBP ranged from 83-92%. One study classified LBP in 160 industrial workers on personal, psychosocial and occupational factors using an Artificial Neural Network (ANN; programs that operate with multiple processing elements or neurons to determine the strength of connections between nodes) with 92% accuracy.25 The next largest study was one in 34,589 people and showed an ANN work on lifestyle and psychosocial characteristics classified LBP with an area under the curve of 0.75. Eleven studies looked at the classification of individuals with chronic LBP.19,20,24,28,29,37,42,54,57,62,64 The sample size of studies in chronic LBP classification ranged from 24 to 171 individuals.19,20,24,28,29,37,42,54,57,62,64 Nine of these studies used input parameters which focused on electromyography and trunk motion data.20,24,28,29,37,42,54,57,62 The accuracy of the machine learning models for CLBP classification ranged from 70-100%.19,20,24,28,29,37,42,54,57,62,64

No studies have used AI/ML techniques to assess LBP prognosis of pre-defined sub-groups on pain and disability outcomes. However, nine studies assessed the prognosis of LBP based on input parameters.21,22,27,30,31,46,51,52,59 Studies examined prognosis prediction using AI/ML techniques of: satisfaction after lumbar stenosis surgery,21 recurrent lumbar disc herniation,22 recovery from acute LBP,27,30 recovery from CLBP,31 poor outcomes following lumbar surgery,46,51 successful outcomes from cognitive behavioural therapy52 and recovery based on pain chart measurements.59 Sample sizes ranged from 71 to 4,665 people. Six studies showed an accuracy of 61‑98%,21,22,27,31,51,52 while three did not report accuracy directly.59,67,68 One study reported an area under the curve of 0.75,30 while the other study reported a sensitivity and specificity of 88% and 86%, respectively.46

Four studies38,48,65,66 assessed the ability of AI/ML approaches to, using existing datasets, diagnose nerve root compression, ‘simple’ LBP, spinal pathology and abnormal illness behaviour in LBP. There models achieved an accuracy of 82% and 90% respectively.38,48,65,66 Two studies aimed to predict vertebral pathologies with an accuracy of 90-92%.58,61 Lastly, one study used a decision support system for LBP diagnosis with an accuracy of 73%.60

No prospective clinical trials have been performed using AI/ML tools for LBP treatment allocation. However, two studies26,43 looked at treatment allocation pathways. One study looked at computer assisted prediction of LBP treatment, but did not report any accuracy values nor clearly the number of treatment pathways.26 The other study used 1288 fictional cases to train the data set and a training sample of 45 humans.43 The highest accuracy for predicting appropriate treatment allocation reported was 72%.43

Five studies35,36,39,45,56 did not clearly fit classification, diagnosis, prognosis or treatment allocation titles. Two studies assessed prediction of pain intensity in LBP based on pain intensity and skin resistance45 and spinal motion data.56 The use of sleep actigraphy to determine daytime pain was assessed in one study using an ANN.36 Another was used to predict neural adaptions based on psychosocial constructs using a Multivariate Pattern analysis.39 Lastly, one study assessed self-report and objective activity data to categorise acute and chronic LBP using an ANN.35

An overview of risk of bias from the NOS is in Table 2. Overall, 29 studies20,23-25,28,29,32,34,38,40-42,44,45,47-50,53-55,57,58,61-66 were case‑control while eight21,22,27,30,31,37,46,52 were cohort studies. Eleven studies did not fit the criteria for case‑control or cohort studies and did not undergo risk of bias assessment.19,26,33,35,36,39,43,51,56,59,60 Of the case-control studies, eight were considered ‘fair’ quality,20,48,55,57,61,64-66 while the other 21 were ‘poor’ quality.23-25,28,29,32,34,38,40-42,44,45,47,49,50,53,54,58,62,63 All eight cohort studies were considered as ‘fair’ quality.21,22,27,30,31,37,46,52

**STarT Back tool**

Overall, 46 studies were included within the STarT Back review (Supplementary Figure 1).13-15,69-111

Reliability and validity are summarised in Supplementary Table 1. Nine studies assessed the internal consistency of the tool, with a Cronbach’s α ranging from 0·51-0·93 (poor to strong)69,76,83,89,99,100,102,104,110. Only one study achieved an internal consistency above 0·9 (strong), which is recommended for use in individuals.102 Nine studies also assessed the test-retest reliability of the STarT Back with the intraclass correlation coefficient and kappa values ranging from 0·65-0·93 (moderate to excellent).75,76,83,88,99,100,102,104,110 Construct validity was assessed in ten studies with correlation values ranging from 0·18-0·75 (weak to strong), however most comparisons were of moderate strength.69,72,75,76,80,83,88,99,104,110 Lastly, the discriminative validity was assessed in eight studies with the area under the curve ranging from 0·65-0·94 (poor to excellent).13,14,69,70,74,83,89,101

For prognosis, STarT Back classification for improving pain or disability are in Supplementary Table 2. Of these, 17 studies assessed pain and disability prognosis with univariate models.71,75,78,81,82,85-87,90,95,97,98,105-109 Of the univariate analyses, eight showed significant prognostic benefits for pain intensity,75,84,86,90,94,98,107,108 13 showed significant prognostic benefits for disability,75,84-87,90,94,95,97,98,103,106,109 while two showed significant prognostic benefits on mixed pain intensity and disability analyses.81,82 Of the multivariate models, two studies showed the STarT Back to predict prognosis for pain intensity adjusted for baseline pain,91,92 while four showed no significant association.72,73,79,94 Eight studies assessed prognosis for disability in multivariate models adjusted for baseline levels of disability with, six studies in favour72,73,84,91,94,103 and two against79,92 a significant association.

Four clinical trials assessed the STarT Back for classification and treatment allocation compared outcomes to standard care (Supplementary Table 3).15,77,96,111 Of these, two were a non‑randomised trials, one which showed significant benefits of stratified care for pain and disability outcomes,96 while the other only showed significant benefits for disability.111 The two RCTs showed no significant effects of stratified care on pain intensity,15,77 while one showed a significant effect for disability.15 One RCT15 and one non-randomised trial111 assessed the cost effectiveness of stratified care when compared with standard care, with no significant differences observed.

**McKenzie method**

Overall, 29 studies were included within the McKenzie review (Supplementary Figure 2).112-140

Eight studies looked at the inter-tester reliability and classification ability of the McKenzie method (Supplementary Table 4).114,116,122,123,132-134,137 Overall, seven studies assessed the reliability with a Kappa value range of 0·02-1·00.114,122,123,132-134,137 Only two of these studies had Kappa ranges greater than 0·6, thus five studies had poor to moderate agreement.141One study also showed that 31% of individuals were not able to be classified with the McKenzie method.116 Validity of the McKenzie method as a classification system cannot be tested, as there is no gold standard comparator.142

Prognosis on pain intensity or disability based on McKenzie principles, such as directional preference, centralisation versus peripheralization and pain pattern classification was assessed in 11 studies (Supplementary Table 5).115,118,121,125,129,131,135,136,138-140 The duration of follow-up of these studies ranged from two weeks to one-year. Four studies reported the follow-up as when the patient was discharged, however did not provide a timeframe.115,131,139,140 Three studies showed that classification was a significant predictor of pain intensity in univariate models,115,136,140 while one did not.118 No studies aimed to assess classification on pain intensity in a multivariate model when adjusted for baseline values. For disability, five studies showed no significant benefit of classification on prognosis,118,129,131,135,138 while five showed a significant effect.115,121,125,139,140 Only two studies assessed disability prognosis this within multivariate models, with one showing significant139 and one non-significant results.138

The search identified 11 clinical trials which used the McKenzie assessment and then provided treatment based on the individuals classification compared to another intervention or treatment (Supplementary Table 6).112,113,117,119,120,124,126-128,130,131 The comparators in the trials consisted of standard physiotherapy,112 chiropractic treatment,113 back-care booklet,113 back school,117 motor control exercise,119,127 endurance exercises,120 first-line care,126 manual therapy,128 general advice,128 intensive strengthening130 and spinal manipulation therapy.131 Five of 11 trials showed significant benefits for pain intensity which favoured McKenzie treatment at the end of intervention.112,113,120,124,126 For disability, four of 11 studies showed significant benefits favouring McKenzie treatment at the end of intervention.112,117,120,124 Three studies112,124,126 assessed McKenzie compared to standard care, with all studies showing significant results favouring McKenzie for pain intensity and two for disability.112,124 Three studies113,120,128 assessed McKenzie compared to advice or education, with two showing significant improvements in pain intensity113,120 and one in disability,120 favouring McKenzie. Compared to passive treatments, such as manual therapy or mobilisations, three studies showed no significant differences for pain intensity and disability.113,128,131 Three studies compared McKenzie to active treatments, with no significant results for pain intensity or disability observed.119,127,130 One study compared McKenzie to Back School, with significant results favouring McKenzie for disability but not pain intensity.117 One study assessed costs with no differences observed between McKenzie therapy and standard chiropractic treatment.113

**DISCUSSION**

AI/ML are becoming more widely used in disease management and has potential to impact LBP treatment.12 This systematic review assessed the current status of these approaches in the management LBP. In comparison to other classification approaches, applying methods of AI/ML for LBP is currently in its infancy. The results of our review show that machine learning tools, such as artificial neural networks and support vector machines, have attempted binary classification (presence of LBP or not), recovery prediction and treatment allocation in LBP. The accuracy of models included in this study ranged from 61‑100%. However, there are several important limitations in existing AI/ML research.

Study sample sizes used for AI/ML based LBP classification or prognosis were typically small for machine learning approaches, with 23 of 48 studies having a sample size less than 100, 22 of 48 studies with a sample size between 100 and 1,000 and only 3 of 48 studies with a sample size greater than 1000. Additionally, 19 of 48 studies typically used a small range of parameters (≤5 factors). This may be a limitation given most AI/ML studies of non-specific LBP aimed to classify individuals using only physical factors, such as trunk range of motion, electromyography and sitting posture;20,23,24,28,29,32,37,40-42,54,57 omitting important psychosocial parameters that are known to be involved in patients with LBP. Only Darvishi et al.25 and Parsaeian et al.44 utilised a range of physical, psychological and social factors for the classification of LBP, however did not attempt sub-classification that delineate sub-groups that could benefit from specific treatments. LBP sub-classification is important as LBP, especially chronic (>12wks) LBP, as it is characterised by changes to a series of systems: biological, psychosocial and the central nervous systems and there are likely sub-groups within this population.143 Notably, some studies applied many models to small CLBP datasets (n<100) to yield highly accurate results, however these were only focused on the binary classification, determining only the presence of CLBP.20,24,28,29,42 In machine learning, normally, the sample size should be no less than 2k cases (where k is the number of features), with a preference of 5 x 2k.144 Therefore these studies may be prone to overfitting of data and the best fit model is likely not applicable to other LBP samples.145 Overall, 25 studies within this review assessed the role of machine learning on classification of individuals with LBP. To develop a robust sub-classification tool, various conditions such as reliability, validity, accuracy, ease of implementation, treatment allocation yielding clinically meaningful benefits and reductions in healthcare costs should be met.146 The current evidence for the use of AI/ML highlights that the utility of these approaches is yet to be realised in a clinically meaningful way.

For comparison, we also conducted systematic reviews of two other classification systems for back pain: STarT Back tool (classifies people in to low, medium and high risk of developing chronic pain based on physical and psychosocial factors)13 and the McKenzie method (diagnosing movement preferences; e.g. spinal extension versus flexion).16 The reliability (i.e. the consistency of the classification system over repeated attempts with the same patient)147 of the McKenzie method was poor to moderate114,116,122,123,132-134,137 and moderate to excellent for the STarT Back tool.75,76,83,88,99,100,102,104,110 This limits the ability of the McKenzie method to be a useful classification system for people with LBP, as this impacts the ability to identify a movement or structure that benefits from a specific treatment.142 Construct validity (i.e. degree of which the measure reflects what it is trying to attain)147 of the STarT Back tool ranged from weak to strong69,72,75,76,80,83,88,99,104,110 and discriminative validity (i.e. the ability to discriminate between various groups of individuals or sub-groups)147 was poor to excellent.13,14,69,70,74,83,89,101 Three studies achieved poor discriminative validity for a singular subscale,14,89,101 while all other values were above acceptable. Validity of the McKenzie method as a classification system has not and cannot be assessed, as there is no gold standard comparator.142 Based on our findings from these two systematic reviews, if AI/ML is to make an impact on LBP management, it will likely need to develop greater reliability and validity compared to current approaches and advance sub-groups to improve clinical and societal outcomes through appropriate treatment allocation (Table 3).

In assessing the ability of a classification system to predict prognosis (i.e. the trajectory of a condition based on certain sub-group factors) of people with LBP, it is critical to account for the patients' pain and disability when they are first assessed, as these factors are the strongest and most consistent predictors of pain and disability in the months after LBP incidence.148-151 The STarT Back tool was typically (in six72,73,84,91,94,103 of eight79,92 studies and 2,080 of 2,634 patients) able to predict future disability, but this was less consistent for pain intensity (two91,92 of six72,73,79,94 studies and 348 of 1,899 patients). For the McKenzie method, no studies assessed the effectiveness of the classification method on future pain intensity whilst accounting for baseline values. For disability, two studies of McKenzie assessed disability prognosis this within multivariate models, with results mixed (significant in one of two studies and 109 of 832 patients).138,139 The utility of the tool to effect overall improvements in patient outcomes has not been tested extensively for the STarT Back tool. One non-randomised trial showed significant benefits for pain intensity and disability when implementing the STarT Back compared to usual case (n=582).96 Of the two RCTs, neither showed benefits of stratification on pain intensity (1,324 patients), however one showed significant improvement for disability compared to usual care (one of two studies and 568 of 1,324 patients).15,77 The McKenzie method has been tested in 11 RCTs,112,113,117,119,120,124,126-128,130,131 but in comparison to other active and passive treatment approaches is not more effective.

To build on current machine learning approaches, research should investigate the ability to create sub-groups of individuals with LBP that considers a broader range of biopsychosocial factors similar to that of the STarT back tool. The use of a broader range of clinical factors incorporated within an AI/ML approach using a large training data set may enable for more reliability, validity, prognostic capacity, and improved stratification of treatment for patients with LBP.9 Such an approach may therefore lead to improved clinical outcomes for clients and reduce healthcare expenditure, however this is yet to be determined. To date, only one study has aimed to employ this approach in LBP with a narrow set of physical factors.43 Oude et al.43 used 1,288 fictional cases to develop a model of self-referral in LBP, which was then applied to 45 real cases with a modest accuracy of 72%. Furthermore, the study did not assess if the model could lead to improved clinical outcomes and reduced healthcare costs.43 A limitation of such approaches is that they failed to consider psychosocial and central nervous system factors that are associated with the condition, such as kinesiophobia,152 pain catastrophizing,153 pain beliefs,154 pain self-efficacy,155 depression,5 anxiety,5 occupational factors,156 sensory changes157 and structural and functional changes to the brain.158,159 Including these factors may allow for specific sub-groups to be identified that could benefit from targeted treatments to maximise clinical benefits. Future models that aim to classify treatment approaches need to consider these broader psychosocial and behavioural factors to enhance accuracy and clinical utility of the model.

The strengths of the current study include the use of broad search terms to identify all the relevant literature pertaining to the use of artificial intelligence in LBP. Even with these terms, we were only able to identify 185 articles for title/abstract screening. Furthermore, we completed two additional systematic reviews to contrast how machine learning could build on current classification approaches in LBP. For limitations, for clinical trials, due to the low number of studies and heterogeneity between studies, meta-analysis could not be performed. Furthermore, we considered the overall interaction of STarT Back classification tool (e.g. combination of all groups) when assessing the effectiveness for the intervention on pain, disability and costs. Some groups may have had significant effects, while others did not.15 However, it is important to determine if we can develop a tool where all sub-groups benefit from specific treatments. Overall, we provide a clear summary of what the benefits of McKenzie and STarT Back could be.

Machine learning has the potential to improve the management of LBP via sub-classification of an otherwise homogenous diagnosis such as non-specific LBP. Identifying relevant sub-groups among patients with LBP would permit the determination of diagnostic categories that inform clinical decision-making and treatment choice. This systematic review found that current machine learning approaches are reported to have high accuracy, however, are often applied to small data sets with multiple models. To determine the utility of such approaches in future research, studies implementing machine learning in LBP need examine larger sample sizes, examine a variety of known risk factors across multiple domains (e.g. spinal tissue, psychosocial and central nervous system) in each model and attempt sub-classification through data clustering within the model. The classification approaches need to be reliable, robust, evaluated, detect sub-groups with different prognosis and inform allocation of patients to treatment such that patient outcomes and/or healthcare costs are, overall, improved. Ultimately, this kind of approach to sub-classification has the potential to drive improvements in the global health-related burden of disease.

**METHODS**

**Search strategy**

These systematic reviews were prospectively registered with PROSPERO prior to data extraction beginning (as registration numbers are still pending protocols were uploaded to the Open Science Framework: AI/ML <https://osf.io/a8nzt/> STarT Back and McKenzie <https://osf.io/ztehm/>). Six databases were searched to September 2019 with the following limits: MEDLINE (Nil), CINAHL (Exclude MEDLINE), SPORTDiscus (Nil), EMBASE (Exclude MEDLINE), PsycINFO and CENTRAL (Exclude MEDLINE and EMBASE). For the machine learning systematic review, IEEE Xplore (Nil) was also searched. Search strategy (1) included MeSH terms for ‘low back pain’ AND ‘artificial intelligence’ (Supplementary Table 7), (2) searches included MeSH terms for ‘low back pain’, and ‘STarT Back Screen’ OR ‘STarT Back Tool’ (Supplementary Table 8) and (3) searches included MeSH terms for ‘low back pain’ and ‘McKenzie’ (Supplementary Table 9). Additional references were searched for through GoogleScholar. Two independent assessors screened the studies and extracted the data for machine learning (SDT and DLB), the STarT Back tool (SDT and DLB) and the McKenzie method (SDT and XZ). All disagreements were addressed via an adjudicator (PJO).

**Inclusion and exclusion criteria**

For inclusion, studies must have examined LBP and the utilisation of AI/ML techniques, the STarT Back or McKenzie method in humans. LBP was defined as pain localized below the costal margin and above the inferior gluteal folds.160 No restrictions were included based on race, sex or age. Studies were required to be a full peer reviewed journal or full conference publication (i.e. grey literature excluded). For AI/ML approaches in LBP, there was no restriction on study design, to ensure all research on this approach to date was identified. For STarT Back or McKenzie there was the inclusion criterion that the study must have examined (a) reliability, (b) validity, (c) prognosis and/or (d) treatment effects (such as in a clinical trial). There was no restriction on study design as long as those topics were addressed. Exclusion criteria were: not peer reviewed or full conference abstract, not English language, not low back pain, not AI/ML or STarT Back or McKenzie classification (e.g. if not clear individuals were assessed and treated via their profile) and not original research. AI/ML studies that did not evaluate the role of AI/ML in patient classification, prognosis or treatment (e.g. automated radiographic image analysis, automated pain diagram analysis) were excluded.

**Data extraction**

Data extracted included relevant publication information (i.e. author, title, year, journal), study design (e.g. cross sectional), study overview (free text), number of participants, type of LBP (e.g. acute, subacute, chronic, unclear) and summary of authors conclusions (free text). For A/ML articles further extraction acquired the AI/ML techniques implemented, parameters used as inputs, whether data was split into training and testing datasets and the main results (e.g. the highest sensitivity, specificity, accuracy and area under the curve that are available). For both the STarT Back and McKenzie reviews, additional data was extracted for reliability, validity, prognosis and treatment effects from sub-classification (e.g. significant improvements to pain intensity, disability and healthcare costs). When it is was not possible to extract the required data, this information was requested from the authors a minimum of three times over a four-week period. Any discrepancies were discussed by the two independent assessors with disagreements addressed via an adjudicator (PJO).

**Definitions used in the systematic review**

For studies of AI/ML in LBP, we considered the following categories of classification, sub-classification, prognosis, diagnosis and treatment allocation. Classification was considered as the ability to discriminate individuals with LBP from healthy populations, while sub-classification was defined as the ability to sub-group individuals with LBP based on different clinical characteristics (e.g. anatomical, psychological and nervous system alterations).146 Prognosis was considered the ability of clinical variables or a assessed sub-group to predict recovery or non-recovery (i.e. clinical course) of pain intensity or disability from LBP.161 Diagnosis was defined as the ability to determine the cause of LBP, which could be based on anatomical, psychological and nervous system factors.162 Treatment allocation was determined to be the prediction of a type of treatment that could benefit a certain individual with LBP.163 Studies that did not clearly fit in these definitions were classed as ‘other’ studies.

**Cut-offs for reliability and validity**

Internal consistency (i.e. the degree of which components of a measure are related) was considered acceptable if Cronbach’s α values ranged from 0·7 to 0·9, while values ≥0·9 were considered strong.147 Test-retest (i.e. the consistency of the classification system over repeated attempts with the same patient) was considered as acceptable above an intraclass correlation coefficient (ICC) of ≥0·7, whereas values ≥0·9 are considered acceptable for individuals, therefore we considered these values as strong.147,164 When Kappa scores for intra-rater (i.e. agreement of repeated measurements on the same patient) or inter-tester (i.e. the agreement of measurements between different clinicians) reliability were available, values were considered as poor agreement (0-0·2), slight agreement (0·21-0·40), moderate agreement (0·41-0·6), good agreement (0·61-0·8) and excellent agreement (0·81-1).123 As recommended for disability research, construct validity correlations (i.e. degree of which the measure reflects what it is trying to attain)147 above 0·6 were considered as strong, 0·3-0·6 as moderate, and below 0·3 as weak.147,165 Discriminative validity (i.e. the ability to discriminate between various groups of individuals or sub-groups)147 followed principles set by Hill and colleagues13 for the STarT Back with an area under the curve of 0·7-<0·8 indicating acceptable discrimination, 0·8-<0·9 indicating excellent discrimination and ≥0·9 indicating outstanding discrimination.

**Risk of bias**

Risk of bias was assessed by the Newcastle-Ottawa Scale (NOS: <http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp>), which is recommend for quality assessment of case-control and cohort studies by the Cochrane Collaboration group.166 The NOS is split into selection, comparability and ascertainment of exposure/outcome categories, with a maximum score of nine points awarded. Based on this, studies were determined to be good, fair or poor quality as previously determined.166 The methodological quality was determined by two independent reviewers (SDT and DLB). Results were compared with disagreements discussed to reach a verdict, with adjudication by PJO if necessary.

**Data Availability:** Available upon request.

**Acknowledgements**: Scott Tagliaferri is supported by an Australian Government Research Training Program (RTP) Scholarship.

**Competing interests**: The authors declare no competing interests.

**Funding:** No funding was received for this research.

**Contributors:**

SDT – Conception of review, data extraction (AI/ML, McKenzie and STarT Back), risk of bias assessment, preparation and revision of manuscript.

MAT – Feedback, guidance and revision of manuscript.

XZ – Data extraction (McKenzie), revision of manuscript.

PJO – Feedback, guidance and revision of manuscript.

CTM – Feedback, guidance and revision of manuscript.

TW – Feedback, guidance and revision of manuscript.

DLB – Conception of review, database searches, data extraction (AI/ML and STarT Back), risk of bias assessment and revision of manuscript.

**REFERENCES**

1 Vos, T. *et al.* Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: A systematic analysis for the global burden of disease study 2010. *The Lancet* **380**, 2163-2196 (2012).

2 Walker, B., Muller, R. & Grant, W. Low back pain in Australian adults: The economic burden. *Asia Pacific Journal of Public Health* **15**, 79-87 (2003).

3 Martin, B. I. *et al.* Expenditures and health status among adults with back and neck problems. *JAMA* **299**, 656-664 (2008).

4 Froud, R. *et al.* A systematic review and meta-synthesis of the impact of low back pain on people’s lives. *BMC Musculoskeletal Disorders* **15**, 50 (2014).

5 Stubbs, B. *et al.* The epidemiology of back pain and its relationship with depression, psychosis, anxiety, sleep disturbances, and stress sensitivity: Data from 43 low-and middle-income countries. *General Hospital Psychiatry* **43**, 63-70 (2016).

6 Verbunt, J. A., Smeets, R. J. & Wittink, H. M. Cause or effect? Deconditioning and chronic low back pain. *Pain* **149**, 428-430 (2010).

7 Gatchel, R. J., Peng, Y. B., Peters, M. L., Fuchs, P. N. & Turk, D. C. The biopsychosocial approach to chronic pain: Scientific advances and future directions. *Psychological Bulletin* **133**, 581 (2007).

8 Bardin, L. D., King, P. & Maher, C. G. Diagnostic triage for low back pain: A practical approach for primary care. *Medical Journal of Australia* **206**, 268-273 (2017).

9 Rabey, M. *et al.* Chronic low back pain is highly individualised: Patterns of classification across three unidimensional subgrouping analyses. *Scandinavian Journal of Pain*, 1-11 (2019).

10 Diller, G.-P. *et al.* Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients. *European Heart Journal* **40**, 1069-1077 (2019).

11 Wu, C.-C. *et al.* Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine* **170**, 23-29 (2019).

12 Lötsch, J. & Ultsch, A. Machine learning in pain research. *Pain* **159**, 623 (2018).

13 Hill, J. C. *et al.* A primary care back pain screening tool: Identifying patient subgroups for initial treatment. *Arthritis Care & Research* **59**, 632-641 (2008).

14 Hill, J. C., Dunn, K. M., Main, C. J. & Hay, E. M. Subgrouping low back pain: a comparison of the STarT Back Tool with the Örebro Musculoskeletal Pain Screening Questionnaire. *European Journal of Pain* **14**, 83-89 (2010).

15 Hill, J. C. *et al.* Comparison of stratified primary care management for low back pain with current best practice (STarT Back): A randomised controlled trial. *The Lancet* **378**, 1560-1571 (2011).

16 McKenzie, R. & May, S. *The lumbar spine: mechanical diagnosis & therapy. Vol. 1*. (Spinal Publications New Zealand, 2003).

17 Lam, O. T. *et al.* Effectiveness of the McKenzie method of mechanical diagnosis and therapy for treating low back pain: literature review with meta-analysis. *Journal of Orthopaedic & Sports Physical Therapy* **48**, 476-490 (2018).

18 Almeida, M., Saragiotto, B., Richards, B. & Maher, C. G. Primary care management of non‐specific low back pain: key messages from recent clinical guidelines. *Medical Journal of Australia* **208**, 272-275 (2018).

19 Lee, J. *et al.* Machine learning–based prediction of clinical pain using multimodal neuroimaging and autonomic metrics. *pain* **160**, 550-560 (2019).

20 Ashouri, S. *et al.* A novel approach to spinal 3-D kinematic assessment using inertial sensors: Towards effective quantitative evaluation of low back pain in clinical settings. *Computers in biology and medicine* **89**, 144-149 (2017).

21 Azimi, P., Benzel, E. C., Shahzadi, S., Azhari, S. & Mohammadi, H. R. Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis. *Journal of neurosurgery* **20**, 300-305 (2014).

22 Azimi, P., Mohammadi, H. R., Benzel, E. C., Shahzadi, S. & Azhari, S. Use of artificial neural networks to predict recurrent lumbar disk herniation. *Clinical Spine Surgery* **28**, E161-E165 (2015).

23 Bishop, J. B., Szpalski, M., Ananthraman, S. K., McIntyre, D. R. & Pope, M. H. Classification of low back pain from dynamic motion characteristics using an artificial neural network. *Spine* **22**, 2991-2998 (1997).

24 Caza-Szoka, M., Massicotte, D., Nougarou, F. & Descarreaux, M. in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* 6409-6412 (IEEE).

25 Darvishi, E., Khotanlou, H., Khoubi, J., Giahi, O. & Mahdavi, N. Prediction effects of personal, psychosocial, and occupational risk factors on low back pain severity using artificial neural networks approach in industrial workers. *Journal of Manipulative and Physiological Therapeutics* **40**, 486-493 (2017).

26 Gal, N., Stoicu-Tivadar, V., Andrei, D., Nemeş, D. I. & Nădăşan, E. Computer assisted treatment prediction of low back pain pathologies. *Studies in health technology and informatics* **197**, 47-51 (2014).

27 Hallner, D. & Hasenbring, M. Classification of psychosocial risk factors (yellow flags) for the development of chronic low back and leg pain using artificial neural network. *Neuroscience letters* **361**, 151-154 (2004).

28 Hu, B., Kim, C., Ning, X. & Xu, X. Using a deep learning network to recognise low back pain in static standing. *Ergonomics* **61**, 1374-1381 (2018).

29 Hung, C.-C., Shen, T.-W., Liang, C.-C. & Wu, W.-T. Using surface electromyography (SEMG) to classify low back pain based on lifting capacity evaluation with principal component analysis neural network method. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 18-21 (2014).

30 Jarvik, J. G. *et al.* Long-term outcomes of a large prospective observational cohort of older adults with back pain. *The Spine Journal* **18**, 1540-1551 (2018).

31 Jiang, N., Luk, K. D.-K. & Hu, Y. A machine learning-based surface electromyography topography evaluation for prognostic prediction of functional restoration rehabilitation in chronic low back pain. *Spine* **42**, 1635-1642 (2017).

32 Jin-Heeku. Analysis of sitting posture using wearable sensor data and support vector machine model. *Medico-Legal Update* **1**, 334-338 (2018).

33 Kadhim, M. A. FNDSB: A fuzzy-neuro decision support system for back pain diagnosis. *Cognitive Systems Research* **52**, 691-700 (2018).

34 Le Duff, F. *et al.* Sharing medical data for patient path analysis with data mining method. *Studies in health technology and informatics*, 1364-1368 (2001).

35 Liszka-Hackzell, J. J. & Martin, D. P. Categorization and analysis of pain and activity in patients with low back pain using a neural network technique. *Journal of Medical Systems* **26**, 337-347 (2002).

36 Liszka-Hackzell, J. J. & Martin, D. P. Analysis of nighttime activity and daytime pain in patients with chronic back pain using a self-organizing map neural network. *Journal of Clinical Monitoring and Computing* **19**, 411-414 (2005).

37 Magnusson, M. L. *et al.* Range of motion and motion patterns in patients with low back pain before and after rehabilitation. *Spine* **23**, 2631-2639 (1998).

38 Mathew, B., Norris, D., Hendry, D. & Waddell, G. Artificial intelligence in the diagnosis of low-back pain and sciatica. *Spine* **13**, 168-172 (1988).

39 Meier, M. L. *et al.* Pain-related fear—dissociable neural sources of different fear constructs. *eNeuro* **5** (2018).

40 Riveros, N. A. M., Espitia, B. A. C. & Pico, L. E. A. Comparison between K-means and self-organizing maps algorithms used for diagnosis spinal column patients. *Informatics in Medicine Unlocked* **16**, 100206 (2019).

41 Oliver, C. Artificial intelligence in the detection of low back pain. *Journal of Orthopaedic Rheumatology* **8**, 207-210 (1995).

42 Oliver, C. & Atsma, W. Artificial intelligence analysis of paraspinal power spectra. *Clinical Biomechanics* **11**, 422-424 (1996).

43 Oude Nijeweme-d'Hollosy, W. *et al.* Evaluation of three machine learning models for self-referral decision support on low back pain in primary care. *International Journal of Medical Informatics* **110**, 31-41, doi:10.1016/j.ijmedinf.2017.11.010 (2018).

44 Parsaeian, M., Mohammad, K., Mahmoudi, M. & Zeraati, H. Comparison of logistic regression and artificial neural network in low back pain prediction: second national health survey. *Iranian Journal of Public Health* **41**, 86 (2012).

45 Sari, M., Gulbandilar, E. & Cimbiz, A. Prediction of low back pain with two expert systems. *Journal of Medical Systems* **36**, 1523-1527 (2012).

46 Shamim, M. S., Enam, S. A. & Qidwai, U. Fuzzy Logic in neurosurgery: Predicting poor outcomes after lumbar disk surgery in 501 consecutive patients. *Surgical neurology* **72**, 565-572 (2009).

47 Silva, L. *et al.* Recurrence quantification analysis and support vector machines for golf handicap and low back pain EMG classification. *Journal of Electromyography and Kinesiology* **25**, 637-647 (2015).

48 Vaughn, M. L., Cavill, S. J., Taylor, S. J., Foy, M. A. & Fogg, A. J. Direct explanations for the development and use of a multi-layer perceptron network that classifies low-back-pain patients. *International Journal of Neural Systems* **11**, 335-347, doi:10.1142/S0129065701000801 (2001).

49 Abdullah, A. A., Yaakob, A. & Ibrahim, Z. Prediction of Spinal Abnormalities Using Machine Learning Techniques. *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, 1-6 (2018).

50 Al Imran, A., Rifat, M. R. I. & Mohammad, R. Enhancing the classification performance of lower back pain symptoms using genetic algorithm-based feature selection. *Proceedings of International Joint Conference on Computational Intelligence*, 455-469 (2020).

51 Andrei, D. *et al.* Computer aided patient evaluation in the low back pain pathology. *2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics*, 27-30 (2015).

52 Barons, M. J., Parsons, N., Griffiths, F. & Thorogood, M. A comparison of artificial neural network, latent class analysis and logistic regression for determining which patients benefit from a cognitive behavioural approach to treatment for non-specific low back pain. *2013 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, 7-12 (2013).

53 Bounds, D. G., Lloyd, P. J. & Mathew, B. G. A comparison of neural network and other pattern recognition approaches to the diagnosis of low back disorders. *Neural Networks* **3**, 583-591 (1990).

54 Caza-Szoka, M., Massicotte, D. & Nougarou, F. Naive Bayesian learning for small training samples: Application on chronic low back pain diagnostic with sEMG sensors. *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, 470-475 (2015).

55 Chan, H., Zheng, H., Wang, H., Sterritt, R. & Newell, D. Smart mobile phone based gait assessment of patients with low back pain. *2013 Ninth International Conference on Natural Computation (ICNC)*, 1062-1066 (2013).

56 Dickey, J. P., Pierrynowski, M. R., Galea, V., Bednar, D. A. & Yang, S. X. Relationship between pain and intersegmental spinal motion characteristics in low-back pain subjects. *SMC 2000 conference proceedings.* **1**, 260-264 (2000).

57 Du, W. *et al.* Recognition of chronic low back pain during lumbar spine movements based on surface electromyography signals. *IEEE Access* **6**, 65027-65042 (2018).

58 Karabulut, E. M. & Ibrikci, T. Effective automated prediction of vertebral column pathologies based on logistic model tree with SMOTE preprocessing. *Journal of medical systems* **38**, 50 (2014).

59 Li, H.-X., Wang, Y. & Zhang, G. Probabilistic fuzzy classification for stochastic data. *IEEE Transactions on Fuzzy Systems* **25**, 1391-1402 (2017).

60 Lin, L., Hu, P. J.-H. & Sheng, O. R. L. A decision support system for lower back pain diagnosis: Uncertainty management and clinical evaluations. *Decision Support Systems* **42**, 1152-1169 (2006).

61 Mathew, B., Norris, D., Mackintosh, I. & Waddell, G. Artificial intelligence in the prediction of operative findings in low back surgery. *British journal of neurosurgery* **3**, 161-170 (1989).

62 Olugbade, T. A., Bianchi-Berthouze, N., Marquardt, N. & Williams, A. C. Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 243-249 (2015).

63 Sandag, G. A., Tedry, N. E. & Lolong, S. Classification of lower back pain using K-Nearest Neighbor algorithm. *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 1-5 (2018).

64 Ung, H. *et al.* Multivariate classification of structural MRI data detects chronic low back pain. *Cerebral cortex* **24**, 1037-1044 (2012).

65 Vaughn, M. L., Cavill, S. J., Taylor, S. J., Foy, M. A. & Fogg, A. J. Direct explanations and knowledge extraction from a multilayer perceptron network that performs low back pain classification. *International Workshop on Hybrid Neural Systems*, 270-285 (1998).

66 Vaughn, M., Cavill, S., Taylor, S., Foy, M. & Fogg, A. A full explanation facility for a MLP network that classifies low-back-pain patients. *The Seventh Australian and New Zealand Intelligent Information Systems Conference, 2001* **11**, 335-347 (2001).

67 Jarvik, J. G. *et al.* Long-term outcomes of a large, prospective observational cohort of older adults with back pain. *The Spine Journal: Official Journal of the North American Spine Society* **18**, 1540-1551, doi:10.1016/j.spinee.2018.01.018 (2018).

68 Shamim, M. S., Enam, S. A. & Qidwai, U. Fuzzy Logic in neurosurgery: predicting poor outcomes after lumbar disk surgery in 501 consecutive patients. *Surgical neurology* **72**, 565-572; discussion 572, doi:10.1016/j.surneu.2009.07.012 (2009).

69 Abedi, M. *et al.* Translation and validation of the Persian version of the STarT Back Screening Tool in patients with nonspecific low back pain. *Manual Therapy* **20**, 850-854 (2015).

70 Aebischer, B., Hill, J. C., Hilfiker, R. & Karstens, S. German translation and cross-cultural adaptation of the STarT back screening tool. *PLoS One* **10**, e0132068 (2015).

71 Azevedo, D. C. *et al.* Baseline characteristics did not identify people with low back pain who respond best to a Movement System Impairment-Based classification treatment. *Brazilian Journal of Physical Therapy* (2019).

72 Beneciuk, J. M. *et al.* The STarT back screening tool and individual psychological measures: evaluation of prognostic capabilities for low back pain clinical outcomes in outpatient physical therapy settings. *Physical Therapy* **93**, 321-333 (2013).

73 Beneciuk, J. M., Fritz, J. M. & George, S. Z. The STarT Back Screening Tool for prediction of 6-month clinical outcomes: relevance of change patterns in outpatient physical therapy settings. *Journal of Orthopaedic & Sports Physical Therapy* **44**, 656-664 (2014).

74 Beneciuk, J. M., Robinson, M. E. & George, S. Z. Subgrouping for patients with low back pain: a multidimensional approach incorporating cluster analysis and the STarT Back Screening Tool. *The Journal of Pain* **16**, 19-30 (2015).

75 Bier, J. D., Ostelo, R. W., Van Hooff, M. L., Koes, B. W. & Verhagen, A. P. Validity and reproducibility of the STarT Back Tool (Dutch Version) in patients with low back pain in primary care settings. *Physical Therapy* **97**, 561-570 (2017).

76 Bruyere, O. *et al.* Validity and reliability of the French version of the STarT Back screening tool for patients with low back pain. *Spine (Phila Pa 1976)* **39**, E123-128, doi:10.1097/brs.0000000000000062 (2014).

77 Cherkin, D. *et al.* Effect of low back pain risk-stratification strategy on patient outcomes and care processes: the match randomized trial in primary care. *Journal of General Internal Medicine* **33**, 1324-1336 (2018).

78 Field, J. & Newell, D. Relationship between STarT Back Screening Tool and prognosis for low back pain patients receiving spinal manipulative therapy. *Chiropractic & Manual Therapies* **20**, 17 (2012).

79 Friedman, B. W., Conway, J., Campbell, C., Bijur, P. E. & John Gallagher, E. Pain one week after an emergency department visit for acute low back pain is associated with poor three‐month outcomes. *Academic Emergency Medicine* **25**, 1138-1145 (2018).

80 Fuhro, F. F., Fagundes, F. R., Manzoni, A. C., Costa, L. O. & Cabral, C. M. Orebro Musculoskeletal Pain Screening Questionnaire Short-Form and STarT Back Screening Tool: Correlation and Agreement Analysis. *Spine* **41**, E931-936, doi:10.1097/brs.0000000000001415 (2016).

81 George, S. Z. & Beneciuk, J. M. Psychological predictors of recovery from low back pain: a prospective study. *BMC Musculoskeletal Disorders* **16**, 49 (2015).

82 Karran, E. L. *et al.* The value of prognostic screening for patients with low back pain in secondary care. *The Journal of Pain* **18**, 673-686 (2017).

83 Karstens, S. *et al.* Validation of the German version of the STarT-Back Tool (STarT-G): a cohort study with patients from primary care practices. *BMC Musculoskeletal Disorders* **16**, 346 (2015).

84 Karstens, S. *et al.* Prognostic ability of the German version of the STarT Back tool: analysis of 12-month follow-up data from a randomized controlled trial. *BMC Musculoskeletal Disorders* **20**, 94 (2019).

85 Katzan, I. L. *et al.* The use of STarT back screening tool to predict functional disability outcomes in patients receiving physical therapy for low back pain. *The Spine Journal* **19**, 645-654 (2019).

86 Kendell, M. *et al.* The predictive ability of the STarT Back Tool was limited in people with chronic low back pain: a prospective cohort study. *Journal of Physiotherapy* **64**, 107-113 (2018).

87 Kongsted, A., Andersen, C. H., Hansen, M. M. & Hestbaek, L. Prediction of outcome in patients with low back pain–a prospective cohort study comparing clinicians' predictions with those of the Start back tool. *Manual Therapy* **21**, 120-127 (2016).

88 Luan, S. *et al.* Cross-cultural adaptation, reliability, and validity of the Chinese version of the STarT Back Screening Tool in patients with low back pain. *Spine* **39**, E974-979, doi:10.1097/brs.0000000000000413 (2014).

89 Matsudaira, K. *et al.* Psychometric properties of the Japanese version of the STarT back tool in patients with low back pain. *PloS one* **11**, e0152019 (2016).

90 Matsudaira, K. *et al.* The Japanese version of the STarT Back Tool predicts 6-month clinical outcomes of low back pain. *Journal of Orthopaedic Science* **22**, 224-229 (2017).

91 Medeiros, F. C., Costa, L. O. P., Added, M. A. N., Salomão, E. C. & Costa, L. d. C. M. Longitudinal monitoring of patients with chronic low back pain during physical therapy treatment using the STarT back screening tool. *Journal of Orthopaedic & Sports Physical Therapy* **47**, 314-323 (2017).

92 Medeiros, F. C., Costa, L. O. P., Oliveira, I. S., Oshima, R. K. & Costa, L. C. M. The use of STarT BACK Screening Tool in emergency departments for patients with acute low back pain: a prospective inception cohort study. *European Spine Journal* **27**, 2823-2830 (2018).

93 Mehling, W., Avins, A., Acree, M., Carey, T. & Hecht, F. Can a back pain screening tool help classify patients with acute pain into risk levels for chronic pain? *European Journal of Pain* **19**, 439-446 (2015).

94 Morso, L. *et al.* The predictive and external validity of the STarT Back Tool in Danish primary care. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* **22**, 1859-1867, doi:10.1007/s00586-013-2690-z (2013).

95 Morsø, L., Kent, P., Manniche, C. & Albert, H. B. The predictive ability of the STarT Back Screening Tool in a Danish secondary care setting. *European Spine Journal* **23**, 120-128, doi:10.1007/s00586-013-2861-y (2014).

96 Murphy, S. E., Blake, C., Power, C. K. & Fullen, B. M. 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)* **41**, 645-652, doi:10.1097/brs.0000000000001305 (2016).

97 Nielsen, A. M., Hestbaek, L., Vach, W., Kent, P. & Kongsted, A. Latent class analysis derived subgroups of low back pain patients–do they have prognostic capacity? *BMC Musculoskeletal Disorders* **18**, 345 (2017).

98 Pagé, I., Abboud, J., Laurencelle, L. & Descarreaux, M. Chronic low back pain clinical outcomes present higher associations with the STarT Back Screening Tool than with physiologic measures: a 12-month cohort study. *BMC Musculoskeletal Disorders* **16**, 201 (2015).

99 Piironen, S. *et al.* Transcultural adaption and psychometric properties of the STarT Back Screening Tool among Finnish low back pain patients. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* **25**, 287-295, doi:10.1007/s00586-015-3804-6 (2016).

100 Pilz, B. *et al.* The Brazilian version of STarT Back Screening Tool-translation, cross-cultural adaptation and reliability. *Brazilian Journal of Physical Therapy* **18**, 453-461 (2014).

101 Pilz, B. *et al.* Construct and discriminant validity of STarT Back Screening Tool–Brazilian version. *Brazilian Journal of Physical Therapy* **21**, 69-73 (2017).

102 Raimundo, A. M. M. *et al.* Portuguese translation, cross-cultural adaptation and reliability of the questionnaire "Start Back Screening Tool"(SBST). (2017).

103 Riis, A., Rathleff, M. S., Jensen, C. E. & Jensen, M. B. Predictive ability of the start back tool: an ancillary analysis of a low back pain trial from Danish general practice. *BMC Musculoskeletal Disorders* **18**, 360 (2017).

104 Robinson, H. S. & Dagfinrud, H. Reliability and screening ability of the StarT Back screening tool in patients with low back pain in physiotherapy practice, a cohort study. *BMC Musculoskeletal Disorders* **18**, 232 (2017).

105 Storm, L., Rousing, R., Andersen, M. O. & Carreon, L. Y. Usefulness of the STarT Back Screening Tool to predict pain problems after lumbar spine surgery. *Danish Medical Journal* **65** (2018).

106 Suri, P., Delaney, K., Rundell, S. D. & Cherkin, D. C. Predictive validity of the STarT Back tool for risk of persistent disabling back pain in a US primary care setting. *Archives of Physical Medicine and Rehabilitation* **99**, 1533-1539 (2018).

107 Tan, C. I. C. *et al.* Predicting outcomes of acute low back pain patients in emergency department: a prospective observational cohort study. *Medicine* **97** (2018).

108 Toh, I., Chong, H.-C., Suet-Ching Liaw, J. & Pua, Y.-H. Evaluation of the STarT Back screening tool for prediction of low back pain intensity in an outpatient physical therapy setting. *Journal of Orthopaedic & Sports Physical Therapy* **47**, 261-267 (2017).

109 Von Korff, M. *et al.* Comparison of back pain prognostic risk stratification item sets. *The Journal of Pain* **15**, 81-89 (2014).

110 Yelvar, G. D. Y. *et al.* Validity and reliablity of Turkish version of STarT Back Screening Tool. (2019).

111 Foster, N. E. *et al.* Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. *The Annals of Family Medicine* **12**, 102-111 (2014).

112 Bid, D. D. A study on central sensitization in chronic non specific low back pain. *Indian Journal of Physiotherapy & Occupational Therapy*, 160 (2018).

113 Cherkin, D. C., Deyo, R. A., Battié, M., Street, J. & Barlow, W. A comparison of physical therapy, chiropractic manipulation, and provision of an educational booklet for the treatment of patients with low back pain. *New England Journal of Medicine* **339**, 1021-1029 (1998).

114 Donahue, M. S., Riddle, D. L. & Sullivan, M. S. Intertester reliability of a modified version of McKenzie's lateral shift assessments obtained on patients with low back pain. *Physical Therapy* **76**, 706-716 (1996).

115 Edmond, S. L. *et al.* Directional preference, cognitive behavioural interventions, and outcomes among patients with chronic low back pain. *Physiotherapy Research International*, e1773 (2019).

116 Flavell, C. A., Gordon, S. & Marshman, L. Classification characteristics of a chronic low back pain population using a combined McKenzie and patho-anatomical assessment. *Manual Therapy* **26**, 201-207 (2016).

117 Garcia, A. N. *et al.* Effectiveness of back school versus McKenzie exercises in patients with chronic nonspecific low back pain: a randomized controlled trial. *Physical Therapy* **93**, 729-747 (2013).

118 Garcia, A. N., Costa, L. d. C. M., Hancock, M. & Costa, L. O. P. Identifying patients with chronic low back pain who respond best to mechanical diagnosis and therapy: secondary analysis of a randomized controlled trial. *Physical Therapy* **96**, 623-630 (2016).

119 Halliday, M. H. *et al.* A randomized controlled trial comparing the McKenzie method to motor control exercises in people with chronic low back pain and a directional preference. *Journal of Orthopaedic & Sports Physical Therapy* **46**, 514-522 (2016).

120 Johnson, O. E., Adegoke, B. O. & Ogunlade, S. O. Comparison of four physiotherapy regimens in the treatment of long-term mechanical low back pain. *Journal of the Japanese Physical Therapy Association* **13**, 9-16 (2010).

121 Karas, R., McIntosh, G., Hall, H., Wilson, L. & Melles, T. The relationship between nonorganic signs and centralization of symptoms in the prediction of return to work for patients with low back pain. *Physical Therapy* **77**, 354-360 (1997).

122 Kilby, J., Stigant, M. & Roberts, A. The reliability of back pain assessment by physiotherapists, using a ‘McKenzie algorithm’. *Physiotherapy* **76**, 579-583 (1990).

123 Kilpikoski, S. *et al.* Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine* **27**, E207-E214 (2002).

124 Long, A., Donelson, R. & Fung, T. Does it matter which exercise? A randomized control trial of exercise for low back pain. *Spine* **29**, 2593-2602 (2004).

125 Long, A., May, S. & Fung, T. The comparative prognostic value of directional preference and centralization: a useful tool for front-line clinicians? *Journal of Manual & Manipulative Therapy* **16**, 248-254 (2008).

126 Machado, L. A., Maher, C. G., Herbert, R. D., Clare, H. & McAuley, J. H. The effectiveness of the McKenzie method in addition to first-line care for acute low back pain: a randomized controlled trial. *BMC Medicine* **8**, 10 (2010).

127 Miller, E. R., Schenk, R. J., Karnes, J. L. & Rousselle, J. G. A comparison of the McKenzie approach to a specific spine stabilization program for chronic low back pain. *Journal of Manual & Manipulative Therapy* **13**, 103-112 (2005).

128 Paatelma, M. *et al.* Orthopaedic manual therapy, McKenzie method or advice only for low back pain in working adults: a randomized controlled trial with one year follow-up. *Journal of Rehabilitation Medicine* **40**, 858-863 (2008).

129 Petersen, T., Christensen, R. & Juhl, C. Predicting a clinically important outcome in patients with low back pain following McKenzie therapy or spinal manipulation: a stratified analysis in a randomized controlled trial. *BMC Musculoskeletal Disorders* **16**, 74 (2015).

130 Petersen, T., Kryger, P., Ekdahl, C., Olsen, S. & Jacobsen, S. The effect of McKenzie therapy as compared with that of intensive strengthening training for the treatment of patients with subacute or chronic low back pain: a randomized controlled trial. *Spine* **27**, 1702-1709 (2002).

131 Petersen, T. *et al.* The McKenzie method compared with manipulation when used adjunctive to information and advice in low back pain patients presenting with centralization or peripheralization: a randomized controlled trial. *Spine* **36**, 1999-2010 (2011).

132 Razmjou, H., Kramer, J. F. & Yamada, R. Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low back pain. *Journal of Orthopaedic & Sports Physical Therapy* **30**, 368-389 (2000).

133 Riddle, D. L. & Rothstein, J. M. Intertester reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine* **18**, 1333-1344 (1993).

134 Seymour, R., Walsh, T., Blankenberg, C., Pickens, A. & Rush, H. Reliability of detecting a relevant lateral shift in patients with lumbar derangement: a pilot study. *Journal of Manual & Manipulative Therapy* **10**, 129-135 (2002).

135 Sufka, A. *et al.* Centralization of low back pain and perceived functional outcome. *Journal of Orthopaedic & Sports Physical Therapy* **27**, 205-212 (1998).

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

137 Werneke, M. W. *et al.* McKenzie lumbar classification: inter-rater agreement by physical therapists with different levels of formal McKenzie postgraduate training. *Spine* **39**, E182-E190 (2014).

138 Werneke, M. W. *et al.* Effect of adding McKenzie syndrome, centralization, directional preference, and psychosocial classification variables to a risk-adjusted model predicting functional status outcomes for patients with lumbar impairments. *journal of Orthopaedic & Sports Physical Therapy* **46**, 726-741 (2016).

139 Werneke, M. W. *et al.* Directional preference and functional outcomes among subjects classified at high psychosocial risk using STarT. *Physiotherapy Research International* **23**, e1711 (2018).

140 Yarznbowicz, R., Tao, M., Owens, A., Wlodarski, M. & Dolutan, J. Pain pattern classification and directional preference are associated with clinical outcomes for patients with low back pain. *Journal of Manual & Manipulative Therapy* **26**, 18-24 (2018).

141 Viera, A. J. & Garrett, J. M. Understanding interobserver agreement: the kappa statistic. *Family Medicine* **37**, 360-363 (2005).

142 Terwee, C. B. *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology* **60**, 34-42 (2007).

143 Hartvigsen, J. *et al.* What low back pain is and why we need to pay attention. *The Lancet* **6736**, 1-12 (2018).

144 Dolnicar, S. A review of unquestioned standards in using cluster analysis for data-driven market segmentation. (2002).

145 Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079-2107 (2010).

146 Fairbank, J. *et al.* The role of classification of chronic low back pain. *Spine* **36**, S19-S42 (2011).

147 Mollayeva, T. *et al.* The Pittsburgh sleep quality index as a screening tool for sleep dysfunction in clinical and non-clinical samples: A systematic review and meta-analysis. *Sleep Medicine Reviews* **25**, 52-73 (2016).

148 Boonstra, A. M., Reneman, M. F., Waaksma, B. R., Schiphorst Preuper, H. R. & Stewart, R. E. Predictors of multidisciplinary treatment outcome in patients with chronic musculoskeletal pain. *Disability and Rehabilitation* **37**, 1242-1250 (2015).

149 Cecchi, F. *et al.* Predictors of response to exercise therapy for chronic low back pain: result of a prospective study with one year follow-up. *Eur J Phys Rehabil Med* **50**, 143-151 (2014).

150 Steffens, D. *et al.* Prognosis of chronic low back pain in patients presenting to a private community-based group exercise program. *European Spine Journal* **23**, 113-119 (2014).

151 van der Hulst, M., Vollenbroek-Hutten, M. M. & IJzerman, M. J. A systematic review of sociodemographic, physical, and psychological predictors of multidisciplinary rehabilitation—or, back school treatment outcome in patients with chronic low back pain. *Spine* **30**, 813-825 (2005).

152 Chou, R. & Shekelle, P. Will this patient develop persistent disabling low back pain? *JAMA* **303**, 1295-1302 (2010).

153 Picavet, H. S. J. Pain catastrophizing and kinesiophobia: Predictors of chronic low back pain. *American Journal of Epidemiology* **156**, 1028-1034, doi:10.1093/aje/kwf136 (2002).

154 Ng, S. K. *et al.* Negative beliefs about low back pain are associated with persistent high intensity low back pain. *Psychology, Health & Medicine* **22**, 790-799 (2017).

155 Jackson, T., Wang, Y., Wang, Y. & Fan, H. Self-efficacy and chronic pain outcomes: A meta-analytic review. *The Journal of Pain* **15**, 800-814 (2014).

156 Steenstra, I., Verbeek, J., Heymans, M. & Bongers, P. Prognostic factors for duration of sick leave in patients sick listed with acute low back pain: A systematic review of the literature. *Occupational and Environmental Medicine* **62**, 851-860 (2005).

157 den Bandt, H. L. *et al.* Pain mechanisms in low back pain: A systematic review and meta-analysis of mechanical quantitative sensory testing outcomes in people with non-specific low back pain. *Journal of Orthopaedic & Sports Physical Therapy*, 1-55 (2019).

158 Kregel, J. *et al.* Structural and functional brain abnormalities in chronic low back pain: A systematic review. *Seminars in arthritis and rheumatism* **45**, 229-237 (2015).

159 Mansour, A. R. *et al.* Brain white matter structural properties predict transition to chronic pain. *PAIN* **154**, 2160-2168 (2013).

160 Van Tulder, M. *et al.* Chapter 3 European guidelines for the management of acute nonspecific low back pain in primary care. *European Spine Journal* **15**, 169-191 (2006).

161 Hayden, J., Dunn, K., Van der Windt, D. & Shaw, W. What is the prognosis of back pain? *Best Practice & Research Clinical Rheumatology* **24**, 167-179 (2010).

162 Koes, B. W., van Tulder, M. W. & Thomas, S. Diagnosis and treatment of low back pain. *The BMJ* **332**, 1430-1434, doi:10.1136/bmj.332.7555.1430 (2006).

163 Alrwaily, M. *et al.* Treatment-based classification system for low back pain: revision and update. *Physical Therapy* **96**, 1057-1066 (2016).

164 Lohr, K. N. Assessing health status and quality-of-life instruments: Atributes and review criteria. *Quality of Life Research* **11**, 193-205 (2002).

165 Andresen, E. M. Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation* **81**, S15-S20 (2000).

166 Wells, G. *et al.* The Newcastle-Ottawa Scale (NOS) for assessing the quality if nonrandomized studies in meta-analyses. *Epub Available from: URL:* [*http://www.ohri.ca/programs/clinical\_epidemiology/oxford.asp*](http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp) (2016).

**Figure 1.** PRISMA diagram of the systematic review of artificial intelligence/machine learning approaches in low back pain research.