npj | Digital Medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

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

Scott D. Tagliaferri[1]✉, Maia Angelova[2], Xiaohui Zhao[3], Patrick J. Owen[1], Clint T. Miller[1], Tim Wilkin[2] and Daniel L. Belavy[1]✉

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 healthcare costs.

npj Digital Medicine (2020)3:93 ; https://doi.org/10.1038/s41746-020-0303-x

INTRODUCTION

Low-back pain (LBP) is the leading cause of disability worldwide[1] and is associated with annual economic costs up to AU $9.2 billion[2] and US $102 billion[3] in Australia and the United States of America, respectively. In addition to economic burden, multiple individual factors (e.g. loss of social identity[4], distress[5] and physical deconditioning[6]) 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 failure[10,11] 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 healthcare 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, reduce healthcare 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 (McKenzie[16] and STarT Back[13]). 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 symptom 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.

[1]Institute for Physical Activity and Nutrition (IPAN), School of Exercise and Nutrition Sciences, Deakin University, Geelong, VIC, Australia. [2]School of Information Technology, Deakin University, Geelong, VIC, Australia. [3]Xi'an University of Architecture & Technology, Beilin, Xi'an, China. ✉email: scott.tagliaferri@deakin.edu.au; belavy@gmail.com

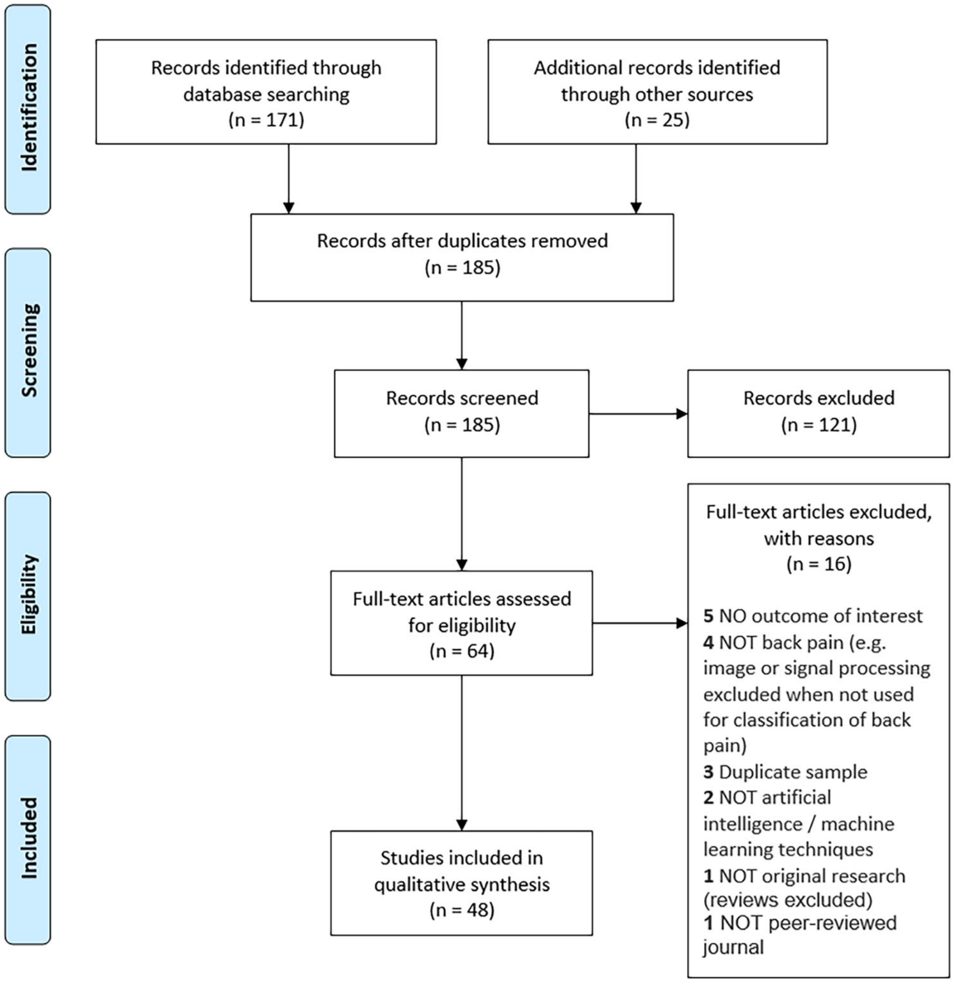
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 171)
Additional records identified through other sources (n = 25)
Records after duplicates removed (n = 185)
Records screened (n = 185)
Records excluded (n = 121)
Full-text articles assessed for eligibility (n = 64)
Full-text articles excluded, with reasons (n = 16)
5 NO outcome of interest
4 NOT back pain (e.g. image or signal processing excluded when not used for classification of back pain)
3 Duplicate sample
2 NOT artificial intelligence / machine learning techniques
1 NOT original research (reviews excluded)
1 NOT peer-reviewed journal
Studies included in qualitative synthesis (n = 48)

Fig. 1 Artificial intelligence PRISMA diagram. Flow of the systematic review of artificial intelligence/machine learning approaches in low-back pain research.

RESULTS

Machine learning

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

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 ranged 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 (multi-layer 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 studies[23,32,40,41] classified LBP and controls based on electromyography, spinal positions and trunk range of motion. Sample sizes of these studies range from 98 to 1510. The accuracy of these studies for classifying LBP ranged from 83 to 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 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 that 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 to 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 therapy[52] and recovery based on pain chart

Table 1. Overview of included studies on machine learning and LBP.

Study	Year	*N*	*N* LBP	*N* CON	Type LBP	AI/ML techniques	Utilised for	Summary	Inputs	Train/Test	Sen	Sp	Acc	AUC	Conclusions
Case–control															
Abdullah et al.[49]	2018	310	210	100	Unclear	K-Nearest Neighbour, Principal Component Analysis, Random Forest	Classification	To predict spinal abnormalities using machine-learning techniques	Pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolisthesis, pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis slope	Yes	—	—	0.85	—	Authors concluded that the KNN classifier outperformed the RF classifier.
Al Imran et al.[50]	2020	310	210	100	Unclear	Random Forest, K-Nearest Neighbour, Support Vector Machine	Classification	Enhancing classification performance in low-back pain symptoms	Pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolisthesis, pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis slope	Yes	—	—	0.92	—	Authors concluded that the application of the genetic algorithm-based feature selection approach can improve classification accuracy.
Ashouri et al.[20]	2017	52	52	28	Chronic	Support Vector Machine	Classification	Spinal 3D kinematic assessment to classify individuals with chronic low-back pain using machine learning	Five trunk flexion and extension parameters	Yes	1.00	1.00	1.00	—	Authors concluded that quantitative techniques provide clinicians and practitioners with improved discriminating means for predicting and diagnosing low-back disorders.
Bishop et al.[23]	1997	183	183	80	Unclear	Artificial Neural Network	Classification	Classifying low-back pain from dynamic motion characteristics	Trunk range of motion and movement velocity	Yes	—	—	0.86	—	Authors concluded a neural network based on kinematic data is an excellent predictive model for the classification of low-back pain.
Bounds et al.[53]	1990	200	200	0	Other	Multi-Layer Perception, K-Nearest Neighbor	Classification	A comparison of neural networks to other pattern recognition approaches for low-back pain	NR	Yes	—	—	0.95	—	Authors concluded that MLP and RBF networks outperform clinicians.
Caza-Szoka et al.[54]	2015	65	43	22	Chronic	Naïve Bayes	Classification	Bayesian learning for electromyography in chronic low-back pain	Electromyography data	Yes	—	—	0.70	—	Authors concluded this paper outlined the advantage of Naïve Bayesian classification models.
Caza-Szoka et al.[24]	2016	24	24	12	Chronic	Artificial Neural Network	Classification	Electromyography array for predicting chronic low-back pain	Electromyography of the paraspinal muscles	Yes	—	—	0.80	—	Authors concluded that a nonlinear analysis can be used for CLBP detection.
Chan et al.[55]	2013	40	20	20	Chronic	Artificial Neural Network, Artificial Neural Network, Multi-Layer Perception, Decision Tree	Classification	A smart phone-based gait assessment to identify people with low-back pain	Gait features	No	—	—	0.88	—	Authors concluded it is feasible to develop a mobile-based tele-care system for monitoring gait.
Darvishi et al.[25]	2017	160	160	92	Unclear	Artificial Neural Network, Logistic Regression, K-Nearest Neighbor	Classification	Prediction of low-back pain severity in industrial workers based on personal, psychological, and occupational factors	Age, gender, body mass index, smoking status, alcohol status, family history, SMWL, job stress, job satisfaction, job security, social relations, force, repetition, posture, and career length	Yes	—	—	0.92	—	Authors concluded that a neural network prediction model was more accurate than regression methods.
Du et al.[57]	2018	171	88	83	Chronic	Support Vector Machine	Classification	Using surface electromyography to detect chronic low-back pain	Electromyography data	Yes	—	—	0.98	—	Authors concluded the models recognised chronic low-back pain with high accuracy.

Table 1 continued

Study	Year	*N*	*N* LBP	*N* CON	Type LBP	AI/ML techniques	Utilised for	Summary	Inputs	Train/Test	Sen	Sp	Acc	AUC	Conclusions
Hu et al.[28]	2018	44	44	22	Chronic	Artificial Neural Network	Classification	Deep learning to identify low-back pain during static standing	Angular rotation, linear translation and centre of pressure measures	Yes	—	—	0.97	0.99	Authors concluded that the deep learning neural networks could be used to accurately differentiate LBP populations from healthy controls using static balance performance.
Hung et al.[29]	2014	52	52	26	Chronic	Artificial Neural Network, Principal Component Analysis	Classification	Electromyography to classify low-back pain from lifting capacity evaluation	Erector spinae muscle activity (including 30 and 50% loading) during lifting tasks	No	0.90	0.88	0.89	0.93	Authors concluded that features with different loadings (including 30 and 50% loading) during lifting can distinguish healthy and back pain subjects.
Jin-Heeku et al.[32]	2018	1510	1510	883	Unclear	Support Vector Machine	Classification	Analysis of sitting posture predicting low-back pain	Data from pressure sensors to assess sitting posture	Yes	1.00	1.00	1.00	—	Authors concluded that a support vector machine can classify individuals with CLBP.
LeDuff et al.[34]	2001	59	59	NR	Unclear	Artificial Neural Network	Classification	Data mining medical records to understand low-back pain treatment pathways	Number of contacts with the different kinds of health professionals, medicines and total costs	Yes	—	—	0.91	—	No specific conclusions.
Melo Riveros et al.[40]	2019	310	310	210	Other	Artificial Neural Network, K-Means Clustering, Self-Organising Map	Classification	Diagnosing spinal pathology from low-back positional characteristics	Pelvic incidence, pelvic inclination, angle of lordosis, sacral slope, pelvic radius and degree of spondylolisthesis	Yes	0.79	0.92	0.83	—	Authors concluded the solution obtained with self-organising maps provides better results with respect to the solution obtained with K-means.
Oliver et al.[41]	1995	98	98	62	Other	Artificial Neural Network	Classification	Electromyography to predict low-back pain.	Electromyography data (power spectra)	Yes	0.82	0.91	0.92	—	Authors concluded that the electromyography signals and ML techniques may be useful for identifying back pain patients.
Oliver et al.[42]	1996	60	60	27	Chronic	Artificial Neural Network	Classification	Electromyography to predict low-back pain	Electromyography data (power spectra)	Yes	0.80	0.79	—	—	Authors stated that artificial intelligence neural networks appear to be a useful method of differentiating paraspinal power spectra in back pain sufferers.
Olugbade et al.[62]	2015	53	23	30	Chronic	Support Vector Machine	Classification	Pain level prediction and classification using kinematics and muscle activity	Trunk flexion kinematics and EMG, sit-to-stand kinematics and EMG and depression	Yes	—	—	0.94	—	Authors concluded the model had very good performance due to thorough analyses.
Parsaeian et al.[44]	2012	34,589	34,589	7286	Unclear	Artificial Neural Network	Classification	Predicting low-back pain based on lifestyle and psychosocial characteristics	Age, sex, education level, urban versus rural, smoker versus non-smoker, strenuous versus non-strenuous working conditions, BMI, mental health disorders and marital status	Yes	—	—	—	0.75	Authors concluded that an artificial neural network approach yielded better performance than logistic regression but that the difference would not be clinically significant.
Sandag et al.[63]	2018	310	210	100	Unclear	K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forest, Decision Tree	Classification	Classification of low-back pain using K-Nearest Neighbour algorithm	Pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolisthesis, pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis slope	Yes	—	—	0.92	—	Authors concluded K-Nearest Neighbour approaches could be used to help further classify low-back pain individuals.

Table 1 continued

Study	Year	*N*	*N* LBP	*N* CON	Type LBP	AI/ML techniques	Utilised for	Summary	Inputs	Train/Test	Sen	Sp	Acc	AUC	Conclusions
Silva et al.[47]	2015	12	12	5	Unclear	Support Vector Machine	Classification	Identifying low-back pain in golfers off muscle activity and swing kinematics	Electromyography during golf swing and kinematic variables of golf swing	Yes	—	—	1.00	—	Authors concluded that low-back pain golfers showed different neuromuscular coordination strategies when compared with asymptomatic golfers.
Ung et al.[64]	2014	94	47	47	Chronic	Support Vector Machine	Classification	Multivariate classification of chronic low-back pain on structural MRI data	Structural brain MRI data	Yes	—	—	0.76	—	Authors concluded support vector machines could classify chronic low-back pain based on grey matter changes.
Karabulut et al.[58]	2014	310	210	100	Unclear	Synthetic Minority Technique, Logistic Model Tree	Diagnosis	Automated predictions of vertebral pathologies with a logistic model tree	Pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolisthesis, pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis slope	Yes	—	—	0.90	—	Authors concluded that the machine-learning techniques reasonably accurate classification.
Mathew et al.[38]	1988	200	200	200	Other	Fuzzy Logic	Diagnosis	Classifying nerve root compression, simple low-back pain, spinal pathology and abnormal illness behaviour.	Age, sex, site of pain, duration of pain, type of onset, relationship to physical activity and movement, neurological symptoms, inappropriate symptoms, red- and yellow-flags in history and spinal deformity	Yes	—	—	0.90	—	Authors stated that the AI techniques can be used for the differential diagnosis of low-back disorders and can outperform clinicians.
Mathew et al.[61]	1989	150	150	0	Surgery	Computer Diagnostic System	Diagnosis	Prediction of operative findings in low-back surgery	Age, sex, site of pain, duration of pain, type of onset, relationship to physical activity and movement, neurological symptoms, inappropriate symptoms, red- and yellow-flags in history and spinal deformity	Yes	—	—	0.92	—	Authors concluded that this computer system has the potential to facilitate assessment on a large number of patients.
Vaughn et al.[65]	1998	198	198	0	Other	Multi-Layer Perception	Diagnosis	Knowledge extraction from a multilayer network for low-back classification	Demographic data, present and past symptoms, pain description/behaviour, finding from physical examination (lumbar spinal movements, tension tests, neurological tests), Oswestry Disability Index, Zung depression index, modified somatic perception questionnaire, the distress and risk assessment method	Yes	—	—	0.96	—	Authors concluded that future work should seek to automatically endure a valid rule for each input case to enhance the network.
Vaughn et al.[66]	2001	196	196	0	Other	Multi-Layer Perception	Diagnosis	MLP network for the classification of low-back pain	Demographic data, present and past symptoms, pain description/behaviour, finding from physical examination (lumbar spinal movements, tension tests, neurological tests), Oswestry Disability Index, Zung depression index, modified somatic perception questionnaire, the distress and risk assessment method	Yes	—	—	0.77	—	Authors concluded a full explanation facility interprets the output on a case-by-case basis.
Vaughn et al.[48]	2001	198	198	198	Other	Artificial Neural Network	Diagnosis	Classifying nerve root compression, simple low-back pain, spinal pathology and abnormal illness behaviour	Demographic data, present and past symptoms, pain description/behaviour, finding from physical examination (lumbar spinal movements, tension tests, neurological tests), Oswestry Disability	Yes	—	—	0.82	—	Authors stated that application of the method leads to the discovery of a number of mis-diagnosed training and test cases and to the development of a more

Table 1 continued

Study	Year	*N*	*N* LBP	*N* CON	Type LBP	AI/ML techniques	Utilised for	Summary	Inputs	Train/Test	Sen	Sp	Acc	AUC	Conclusions
									Index, Zung depression index, modified somatic perception questionnaire and the distress and risk assessment method						optimal low-back-pain MLP network.
Sari et al.[45]	2012	169	169	110	Unclear	Artificial Neural Network, Fuzzy Inference System	Other	Predicting low-back pain intensity based on pain intensity and skin resistance	Skin resistance and pain intensity	Yes	—	—	—	—	Authors stated that their designed systems are effective to predict the pain intensity level objectively.
Cohort															
Magnusson et al.[37]	1998	27	27	0	Chronic	Artificial Neural Network	Classification	Range of motion and motion patterns following rehabilitation in low-back pain	Trunk motion data from eight motion tests	Yes	—	—	0.78	—	Authors stated that a neural network based on kinematic variables is an excellent model for classification of low-back-pain dysfunction.
Azimi et al.[21]	2014	168	168	0	Spinal Stenosis	Artificial Neural Network	Prognosis	Predicting surgical satisfaction for lumbar spinal canal stenosis with artificial neural networks	Age, pain intensity, stenosis ratio, walking distance, Japanese Orthopaedic Association score for assessing LBP, and Neurogenic Claudication Outcome Score	Yes	—	0.41	0.97	0.81	Authors concluded that artificial neural network approach more accurate in predicting 2-year post-surgical satisfaction than a logistic regression model.
Azimi et al.[22]	2015	402	402	0	Recurrent	Artificial Neural Network	Prognosis	Predicting recurrent lumbar disc herniation with artificial neural networks	Age, sex, duration of symptoms, smoking status, recurrent LDH, level of herniation, type of herniation, sports activity, occupational lifting, occupational driving, duration of symptoms, visual analogue scale, the Zung Depression Scale, and the Japanese Orthopaedic Association Score	Yes	—	0.46	0.94	0.84	Authors concluded that artificial neural networks can be used to predict recurrence of lumbar disc herniation.
Barons et al.[52]	2013	701	701	0	Unclear	Artificial Neural Network, Latent Class Analysis, Logistic Regression	Prognosis	Determining who benefits from cognitive behavioural therapy	RMDQ, FABQ, PSE, SF-12, HADS	No	—	—	0.61	—	Authors concluded that artificial neural networks would be the best candidate to support treatment allocation.
Hallner et al.[27]	2004	71	71	0	Acute	Artificial Neural Network	Prognosis	Identifying individuals at risk of chronic low-back pain based on yellow-flags	Pain intensity at the beginning of hospitalisation, Beck Depression Inventory and Kiel Pain Inventory	Yes	0.73	0.97	0.83	—	Authors concluded that this model could contribute to the early detection of risk factors for patients with acute low-back pain, and could assist with avoiding chronicity.
Jarvik et al.[30]	2018	4665	4665	0	Acute	LASSO Model	Prognosis	Predicting recovery from acute low-back pain in older adults	Age, gender, race, ethnicity, education, employment status, marital status, smoking status, the duration of current episode of back or leg pain, back-related claim or lawsuit, patient confidence that their back or leg pain would be completely gone or much better in 3 months, baseline pain-related characteristics, baseline psychological distress, baseline falls, BMI, comorbidity score, baseline diagnosis, spine-related interventions and opioid prescriptions	Yes	—	—	—	0.75	Authors concluded that baseline patient factors were more important than early interventions in explaining disability and pain after 2 years.

Table 1 continued

Study	Year	*N*	*N* LBP	*N* CON	Type LBP	AI/ML techniques	Utilised for	Summary	Inputs	Train/Test	Sen	Sp	Acc	AUC	Conclusions
Jiang et al.[31]	2017	78	30	48	Chronic	Support Vector Machine	Prognosis	Electromyography for prediction of recovery following functional restoration	Electromyography during left lateral bending, right lateral bending, left turning, right turning	Yes	1.00	0.94	0.97	0.89	Authors stated that the tools can be used to identify patients who will respond to functional restoration rehabilitation.
Shamim et al.[46]	2009	501	501	0	Surgery	Fuzzy Inference System	Prognosis	Prediction of poor outcomes following lumbar disc surgery	Sex, BMI, occupation, marital status, use of oral corticosteroids, multilevel disease, epidural steroid injection, duration of symptoms, duration of non-operative treatment, extent of changes on MRI, previous spine surgery, emergency versus elective surgery, operative time, intraoperative complications, operating surgeon and post-op complications	No	0.88	0.86	—	—	Authors concluded a fuzzy inference system is a sensitive method of predicting patients who will fail to improve with surgical intervention.
Other															
Kadhim et al.[33]	2018	10	10	0	Unclear	Fuzzy Inference System	Classification	A decision support system for back pain diagnosis	Sex, height, weight, age and a series of clinical symptoms	No	—	—	0.84	—	Author stated that the proposed system can be used by domain experts (physicians) to help enhance decision-making.
Lee et al.[19]	2019	53	53	0	Chronic	Support Vector Machine	Classification	Prediction of clinical pain intensity from functional connectivity and autonomic states	Functional connectivity and heart rate variability	Yes	—	—	0.92	0.97	Authors concluded that a machine-learning approach model identifies putative biomarkers for clinical pain intensity.
Lin et al.[60]	2006	180	180	0	Unclear	Naïve Bayes	Diagnosis	A decision support system for low-back pain diagnosis	Gender, age, current pain symptoms, clinical pain history, pregnancy history, number and tingling	No	—	—	0.73	—	Authors concluded the system provides an easy-to-follow framework for low-back pain.
Andrei et al.[51]	2015	260	260	0	Other	Fuzzy Inference System	Prognosis	Computer-aided patient evaluation of low-back pathology	Pain, calories, flexion, extension, rotation and lateral flexion range of motion	No	—	—	0.98	—	Authors concluded a complex fuzzy system is essential for lumbar spine pathology.
Li et al.[59]	2017	100	100	0	Unclear	Artificial Neural Network, K-Nearest Neighbor, Fuzzy Inference System	Prognosis	Probabilistic Fuzzy classification for Stochastic data	Pain area, height and width of pain area and ratio	No	—	—	NR	—	Authors concluded more information can be extracted from limited samples using a PFC approach.
Dickey et al.[56]	2000	9	9	0	Chronic	Artificial Neural Network	Other	Relationship between pain and spinal motion characteristics in low-back pain	32 spinal motion parameters	No	—	—	0.99	—	Authors concluded they observed clear patterns of segmental spinal motion in low-back pain.
Liszka-Hackzell et al.[35]	2002	40	40	0	Other	Artificial Neural Network	Other	Categorising individuals with low-back pain based on self-report and activity data	Unclear	Yes	—	—	—	—	Authors stated that that neural network techniques can be applied effectively to categorising patients with acute and chronic low-back pain.
Liszka-Hackzell et al.[36]	2005	18	18	0	Chronic	Artificial Neural Network	Other	Analysis of night-time activity and daytime pain in chronic low-back pain	Measures of sleep quality through actigraphy	Yes	—	—	—	—	Authors concluded that daytime pain levels are not correlated with sleep the night before, nor with the night following.

Table 1 continued

Study	Year	*N*	*N* LBP	*N* CON	Type LBP	AI/ML techniques	Utilised for	Summary	Inputs	Train/Test	Sen	Sp	Acc	AUC	Conclusions
Meier et al.[39]	2018	20	20	0	Chronic	Multivariate Patten Analysis	Other	Predicting neural adaptions based on psychosocial constructs	Bilateral fear-related brain regions including the amygdala, hippocampus, thalamus, anterior cingulate, insula, and medial prefrontal, and orbitofrontal cortices	Yes	—	—	—	—	Authors stated the approach might ultimately help to further understand and dissect psychological pain-related fear.
Gal et al.[26]	2015	15	15	0	Unclear	Fuzzy Inference System	Treatment allocation	Computer-assisted prediction of low-back pain treatment	Sex, age, disability level, daily activity expressed in calories and trunk mobility measures	No	—	—	—	—	Authors concluded the system has the ability to identify the correct treatment and can ensure the quality of the treatment.
Oude et al.[43]	2018	45	45	0	Unclear	Boosted Tree, Decision Tree, Random Forest	Treatment allocation	To determine if self-referral is possible in individuals with low-back pain	Age, well-being index, duration of pain, use of analgesics, history of trauma, use of corticosteroids, presence of specific serious disease, weight loss in past month, constant pain, night-time pain, pain with lifting/sneezing/coughing, radiating pain, reduced muscle strength, cauda equina symptoms, referral preference	Yes	—	—	0.72	—	Authors stated that the study showed possibilities of using ML to support patients with LBP in their self-referral process to primary care.

Acc accuracy, *AI* artificial intelligence, *AUC* area under the curve, — not reported, *ML* machine learning, *Other* study design not case control or cohort, *Sen* sensitivity, *Sp* specificity.

measurements[59]. Sample sizes ranged from 71 to 4665 people. Six studies showed an accuracy of 61–98%[21,22,27,31,51,52], while three did not report accuracy directly[46,59,67]. 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 studies[38,48,65,66] assessed the ability of AI/ML approaches to, using existing data sets, diagnose nerve root compression, 'simple' LBP, spinal pathology and abnormal illness behaviour in LBP. These 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 studies[26,43] 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 studies[35,36,39,45,56] did not clearly fit the classification, diagnosis, prognosis or treatment allocation titles. Two studies assessed the prediction of pain intensity in LBP based on pain intensity and skin resistance[45] 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 shown in Table 2. Overall, 29 studies[20,23–25,28,29,32,34,38,40–42,44,45,47–50,53–55,57,58,61–66] were case–control while eight[21,22,27,30,31,37,46,52] were cohort studies. Eleven studies did not fit the criteria for case–control or cohort studies and did not undergo the 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 Fig. 1)[13–15,68–110]. The reasons for exclusion of STarT Back studies at the full-text stage are presented in Supplementary Table 2.

Reliability and validity are summarised in Supplementary Table 3. Nine studies assessed the internal consistency of the tool, with a Cronbach's α ranging from 0.51 to 0.93 (poor to strong)[68,75,82,88,98,99,101,103,109]. Only one study achieved an internal consistency above 0.9 (strong), which is recommended for use in individuals[101]. Nine studies also assessed the test–retest reliability of the STarT Back with the intraclass correlation coefficient and kappa values ranging from 0.65 to 0.93 (moderate to excellent)[74,75,82,87,98,99,101,103,109]. Construct validity was assessed in ten studies with correlation values ranging from 0.18 to 0.75 (weak to strong); however, most comparisons were of moderate strength[68,71,74,75,79,82,87,98,103,109]. Lastly, the discriminative validity was assessed in eight studies with the area under the curve ranging from 0.65 to 0.94 (poor to excellent)[13,14,68,69,73,82,88,100].

For prognosis, STarT Back classification for improving pain or disability is shown in Supplementary Table 4. Of these, 17 studies assessed pain and disability prognosis with univariate models[70,74,77,80,81,84–86,89,94,96,97,104–108]. Of the univariate analyses, eight showed significant prognostic benefits for pain intensity[74,83,85,89,93,97,106,107], 13 showed significant prognostic benefits

Table 2. Risk of bias assessment using the Newcastle-Ottowa Scale.

Study	Selection				Comparability		Exposure			
Case–control	1	2	3	4	5	6	7	8	9	Total
Abdullah et al.[49]	0	0	0	0	0	0	0	0	0	0/9
Al Imran et al.[50]	0	0	0	0	0	0	0	0	0	0/9
Ashouri et al.[20]	1	0	1	1	0	0	1	0	1	5/9
Bishop et al.[23]	0	0	1	1	0	0	0	1	1	4/9
Bounds et al.[53]	0	0	0	1	0	0	1	1	1	4/9
Caza-Szoka et al.[54]	0	0	0	1	0	0	0	1	1	3/9
Caza-Szoka et al.[24]	0	0	0	1	0	0	0	1	1	3/9
Chan et al.[55]	1	1	1	1	0	0	0	1	1	6/9
Darvishi et al.[25]	0	0	1	1	0	0	0	1	1	4/9
Du et al.[57]	1	1	0	1	0	0	1	1	1	6/9
Hu et al.[28]	1	0	0	1	0	0	0	1	1	4/9
Hung et al.[29]	0	0	0	0	0	0	0	0	1	1/9
Jin-Heeku et al.[32]	0	0	0	0	0	0	0	0	0	0/9
LeDuff et al.[34]	0	0	0	0	0	0	0	1	1	2/9
Melo Riveros et al.[40]	0	0	0	0	0	0	1	1	1	3/9
Oliver et al.[41]	0	0	1	1	0	0	0	1	1	4/9
Oliver et al.[42]	0	0	1	1	0	0	0	1	1	4/9
Olugbade et al.[62]	0	0	0	0	0	0	0	0	0	0/9
Parsaeian et al[44].	0	1	0	1	0	0	0	1	1	4/9
Sandag et al.[63]	0	0	0	0	0	0	0	0	0	0/9
Silva et al.[47]	0	0	0	0	0	0	0	1	1	2/9
Ung et al.[64]	1	1	1	1	0	0	0	1	1	6/9
Karabulut et al.[58]	0	0	0	0	0	0	0	0	0	0/9
Mathew et al.[38]	0	0	0	1	0	0	1	1	0	3/9
Mathew et al.[61]	0	1	0	1	0	0	1	1	1	5/9
Vaughn et al.[65]	0	1	0	1	0	0	1	1	1	5/9
Vaughn et al.[66]	0	1	0	1	0	0	1	1	1	5/9
Vaughn et al.[48]	0	1	0	1	0	0	1	1	1	5/9
Sari et al.[45]	0	0	0	0	0	0	0	1	1	2/9
Cohort	Selection				Comparability		Outcome			Total
Magnusson et al.[37]	1	1	1	1	0	0	1	0	1	6/9
Azimi et al.[21]	1	1	1	1	0	0	1	1	0	6/9
Azimi et al.[22]	1	1	1	1	0	0	0	1	0	5/9
Barons et al.[52]	1	1	1	1	0	0	0	1	1	6/9
Hallner et al.[27]	1	1	1	1	0	0	1	1	0	6/9
Jarvik et al.[30]	1	1	1	1	0	0	1	1	1	7/9
Jiang et al.[31]	1	1	0	1	0	0	0	1	1	5/9
Shamim et al.[46]	1	1	1	1	0	0	0	1	0	5/9
Other[a]	Selection				Comparability		Outcome			Total
Kadhim et al.[33]	—	—	—	—	—	—	—	—	—	—
Lee et al.[19]	—	—	—	—	—	—	—	—	—	—
Lin et al.[60]	—	—	—	—	—	—	—	—	—	—
Andrei et al.[51]	—	—	—	—	—	—	—	—	—	—
Li et al.[59]	—	—	—	—	—	—	—	—	—	—
Dickey et al.[56]	—	—	—	—	—	—	—	—	—	—
Liszka-Hackzell et al.[35]	—	—	—	—	—	—	—	—	—	—
Liszka-Hackzell et al.[36]	—	—	—	—	—	—	—	—	—	—
Meier et al.[39]	—	—	—	—	—	—	—	—	—	—
Gal et al.[26]	—	—	—	—	—	—	—	—	—	—
Oude et al.[43]	—	—	—	—	—	—	—	—	—	—

Higher scores indicate better quality.

[a] Neither case–control nor cohort study design.

for disability[74,83–86,89,93,94,96,97,102,105,108], while two showed significant prognostic benefits on mixed pain intensity and disability analyses[80,81]. Of the multivariate models, two studies showed the STarT Back to predict prognosis for pain intensity adjusted for baseline pain[90,91], while four showed no significant association[71,72,78,93]. Eight studies assessed prognosis for disability in multivariate models adjusted for baseline levels of disability with, six studies in favour[71,72,83,90,93,102] and two against[78,91] a significant association.

Four clinical trials assessed the STarT Back for classification and treatment allocation-compared outcomes to standard care (Supplementary Table 5)[15,76,95,110]. Of these, two were non-randomised trials, one which showed significant benefits of stratified care for pain and disability outcomes[95], while the other only showed significant benefits for disability[110]. The two RCTs showed no significant effects of stratified care on pain intensity[15,76], while one showed a significant effect for disability[15]. One RCT[15] and one non-randomised trial[110] 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 Fig. 2)[111–139]. The reasons for exclusion of McKenzie studies at the full-text stage are presented in Supplementary Table 6.

Eight studies looked at the inter-tester reliability and classification ability of the McKenzie method (Supplementary Table 7)[113,115,121,122,131–133,136]. Overall, seven studies assessed the reliability with a Kappa value range of 0.02−1.00[113,121,122,131–133,136]. Only two of these studies had Kappa ranges >0.6; thus, five studies had poor to moderate agreement[140]. One study also showed that 31% of individuals were not able to be classified with the McKenzie method[115]. Validity of the McKenzie method as a classification system cannot be tested, as there is no gold standard comparator[141].

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 8)[114,117,120,124,128,130,134,135,137–139]. The duration of follow-up of these studies ranged from 2 weeks to 1 year. Four studies reported the follow-up as when the patient was discharged; however, they did not provide a timeframe[114,130,138,139]. Three studies showed that classification was a significant predictor of pain intensity in univariate models[114,135,139], while one did not[117]. No studies aimed to assess the 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[117,128,130,134,137], while five showed a significant effect[114,120,124,138,139]. Only two studies assessed disability prognosis within multivariate models, with one showing significant[138] and one non-significant results[137].

The search identified 11 clinical trials that used the McKenzie assessment and then provided treatment based on the individuals classification compared to another intervention or treatment (Supplementary Table 9)[111,112,116,118,119,123,125–127,129,130]. The comparators in the trials consisted of standard physiotherapy[111], chiropractic treatment[112], back-care booklet[112], back school[116], motor control exercise[118,126], endurance exercises[119], first-line care[125], manual therapy[127], general advice[127], intensive strengthening[129] and spinal manipulation therapy[130]. Five of 11 trials showed significant benefits for pain intensity, which favoured McKenzie treatment at the end of intervention[111,112,119,123,125]. For disability, four of 11 studies showed significant benefits favouring McKenzie treatment at the end of intervention[111,116,119,123]. Three studies[111,123,125] assessed McKenzie compared to standard care, with all studies showing significant results favouring McKenzie for pain intensity and two for disability[111,123]. Three studies[112,119,127] assessed McKenzie compared to advice or education, with two showing significant improvements in pain intensity[112,119] and one in disability[119], favouring McKenzie. Compared to passive treatments, such as manual therapy or mobilisations, three studies showed no significant differences for pain intensity and disability[112,127,130]. Three studies compared McKenzie to active treatments, with no significant results for pain intensity or disability observed[118,126,129]. One study compared McKenzie to Back School, with significant results favouring McKenzie for disability but not pain intensity[116]. One study assessed costs with no differences observed between McKenzie therapy and standard chiropractic treatment[112].

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 ANNs 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 to 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 <100, 22 of 48 studies with a sample size between 100 and 1000 and only 3 of 48 studies with a sample size >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, they did not attempt sub-classification that delineate sub-groups that could benefit from specific treatments. LBP sub-classification is important as LBP, especially chronic (>12 weeks) LBP, 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[142]. Notably, some studies applied many models to small CLBP data sets ($n <$ 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 2^k cases (where k is the number of features), with a preference of 5×2^k[143]. Therefore, these studies may be prone to overfitting of data and the best fit model is likely not applicable to other LBP samples[144]. 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[145]. 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)[146] of the McKenzie method was poor to moderate[113,115,121,122,131–133,136] and moderate to excellent for the STarT Back tool[74,75,82,87,98,99,101,103,109]. 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[141]. Construct validity (i.e. degree of which the measure reflects what it is trying to attain)[146] of the STarT Back tool ranged from weak to strong[68,71,74,75,79,82,87,98,103,109] and discriminative validity (i.e. the ability to discriminate between various groups of individuals or sub-groups)[146] was poor to excellent[13,14,68,69,73,82,88,100]. Three studies achieved poor discriminative validity for a singular subscale[14,88,100], 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[141]. 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[147–150]. The STarT Back tool was typically (in six[71,72,83,90,93,102] of eight[78,91] studies and 2080 of 2634 patients) able to predict future disability, but this was less consistent for pain intensity (two[90,91] of six[71,72,78,93] studies and 348 of 1899 patients). For the McKenzie method, no studies assessed the effectiveness of the classification method on future pain intensity while 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)[137,138]. 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 ($n = 582$)[95]. Of the two RCTs, neither showed benefits of stratification on pain intensity (1324 patients); however, one showed significant improvement for disability compared to usual care (one of two studies and 568 of 1324 patients)[15,76]. The McKenzie method has been tested in 11 RCTs[111,112,116,118,119,123,125–127,129,130], 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 reduced 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 1288 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 fail to consider psychosocial and central nervous system factors that are associated with the condition, such as kinesiophobia[151], pain catastrophizing[152], pain beliefs[153], pain self-efficacy[154], depression[5], anxiety[5], occupational factors[155], sensory changes[156] and structural and functional changes to the brain[157,158]. 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, they 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 to 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 beginning data extraction (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 till 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 10), (2) searches included MeSH terms for 'low back pain' and 'STarT Back Screen' OR 'STarT Back Tool' (Supplementary Table 11) and (3) searches included MeSH terms for 'low back pain' and 'McKenzie' (Supplementary Table 12). Additional references were searched for through GoogleScholar. Two independent assessors screened the studies and extracted the data for machine learning (S.D.T. and D.L.B.), the STarT Back tool (S.D.T. and D.L.B.) and the McKenzie method (S.D.T. and X.Z.). All disagreements were addressed via an adjudicator (P.J.O.).

Table 3. The process of development of (sub-)classification tools for LBP using AI/ML compared to the STarT Back and McKenzie.

	Classification accuracy[a]	Internal consistency[b]	Test–retest reliability[c]	Intra- or inter-rater reliability[d]	Construct validity[e]	Discriminative validity[f]	Prognosis: pain[g]	Prognosis: disability[g]	Treatment: pain[h]	Treatment: disability[h]	Treatment: costs[h]
AI/ML	20/25 (80%)	—	—	—	—	—	—	—	—	—	—
STarT Back	NA	6/9 (67%)	9/9 (100%)	—	5/11 (45%)	8/8 (100%)	2/6 (33%)	6/8 (75%)	1/4 (25%)	3/4 (75%)	0/2 (0%)
McKenzie	NA	—	—	4/10 (40%)	—	—	—	1/2 (50%)	5/11 (45%)	4/11 (36%)	0/1 (0%)

Values reported as number and percentage.

AI/ML artificial intelligence and machine learning, — no studies available or unable to be measured, *NA* not assessed in this systematic review.

[a] Number of AI/ML studies reporting ≥80% accuracy of classification into 'low-back pain' versus 'healthy'.

[b] Internal consistency was considered acceptable if Cronbach's α was ≥0.7[146].

[c] Test–retest was considered as acceptable above an intraclass correlation coefficient (ICC) of ≥0.7[146,163].

[d] Kappa scores for intra-rater and inter-tester reliability were considered good ≥0.61[122].

[e] Construct validity ≥0.6 was considered acceptable[146,164].

[f] Discriminative validity ≥0.7 was considered as acceptable discrimination[13].

[g] Prognosis prediction was considered 'adequate' when the classification approach resulted in statistically significant prediction of outcome after adjusting for baseline pain or disability in multivariate models[147–150].

[h] Treatment effect was considered 'adequate' when the classification approach resulted in a statistically significant improved patients outcomes for pain or disability or healthcare costs in randomised or non-randomised clinical trials.

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 localised below the costal margin and above the inferior gluteal folds[159]. 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 AI/ML articles further extraction acquired the AI/ML techniques implemented, parameters used as inputs, whether data were split into training and testing data sets 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 were extracted for reliability, validity, prognosis and treatment effects from sub-classification (e.g. significant improvements to pain intensity, disability and healthcare costs). When it was not possible to extract the required data, this information was requested from the authors a minimum of three times over a 4-week period. Any discrepancies were discussed by the two independent assessors with disagreements addressed via an adjudicator (P.J.O.).

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)[145]. Prognosis was considered the ability of clinical variables or an assessed sub-group to predict recovery or non-recovery (i.e. clinical course) of pain intensity or disability from LBP[160]. Diagnosis was defined as the ability to determine the cause of LBP, which could be based on anatomical, psychological and nervous system factors[161]. Treatment allocation was determined to be the prediction of a type of treatment that could benefit a certain individual with LBP[162]. 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[146]. 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[146,163]. 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)[122]. As recommended for disability research, construct validity correlations (i.e. degree of which the measure reflects what it is trying to attain)[146] above 0.6 were considered as strong, 0.3−0.6 as moderate, and below 0.3 as weak[146,164]. Discriminative validity (i.e. the ability to discriminate between various groups of individuals or sub-groups)[146] followed principles set by Hill et al.[13] 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 recommended for quality assessment of case−control and cohort studies by the Cochrane Collaboration group[165]. 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[165]. The methodological quality was determined by two independent reviewers (S.D.T. and D.L.B.). Results were compared with disagreements discussed to reach a verdict, with adjudication by P.J.O. if necessary.

DATA AVAILABILITY

All data are available upon request.

Received: 22 January 2020; Accepted: 5 June 2020;
Published online: 09 July 2020

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. *Lancet* **380**, 2163–2196 (2012).
2. Walker, B., Muller, R. & Grant, W. Low back pain in Australian adults: the economic burden. *Asia Pac. J. 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 Musculoskelet. Disord.* **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. *Gen. 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. *Psychol. Bull.* **133**, 581 (2007).
8. Bardin, L. D., King, P. & Maher, C. G. Diagnostic triage for low back pain: a practical approach for primary care. *Med. J. Aust.* **206**, 268–273 (2017).
9. Rabey, M. et al. Chronic low back pain is highly individualised: patterns of classification across three unidimensional subgrouping analyses. *Scand. J. Pain* **19**, 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. *Eur. Heart J.* **40**, 1069–1077 (2019).
11. Wu, C.-C. et al. Prediction of fatty liver disease using machine learning algorithms. *Comput. Meth. Prog. Biomed.* **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 Res.* **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. *Eur. J. 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. *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. *J. Orthop. Sports Phys. Ther.* **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. *Med. J. Aust.* **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. *Comput. Biol. Med.* **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. *J. Neurosurg.* **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. *Clin. Spine Surg.* **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. Surrogate analysis of fractal dimensions from SEMG sensor array as a predictor of chronic low back pain. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* 6409–6412 (IEEE, 2016).
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. *J. Manipulative Physiol. Ther.* **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. *Stud. Health Technol. Inform.* **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. *Neurosci. Lett.* **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. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* 18–21 (IEEE, 2014).
30. Jarvik, J. G. et al. Long-term outcomes of a large prospective observational cohort of older adults with back pain. *Spine J.* **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. *Med.-Leg. Update* **1**, 334–338 (2018).
33. Kadhim, M. A. FNDSB: a fuzzy-neuro decision support system for back pain diagnosis. *Cogn. Syst. Res.* **52**, 691–700 (2018).
34. Le Duff, F. et al. Sharing medical data for patient path analysis with data mining method. *Stud. Health Technol. Informatics.* **84**, 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. *J. Med. Syst.* **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. *J. Clin. Monit. Comput.* **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**, 1–15 (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. *Inform. Med. Unlocked* **16**, 100206 (2019).
41. Oliver, C. Artificial intelligence in the detection of low back pain. *J. Orthop. Rheumatol.* **8**, 207–210 (1995).
42. Oliver, C. & Atsma, W. Artificial intelligence analysis of paraspinal power spectra. *Clin. Biomech.* **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. *Int. J. Med. Inform.* **110**, 31–41 (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. *Iran. J. Public Health* **41**, 86 (2012).
45. Sari, M., Gulbandilar, E. & Cimbiz, A. Prediction of low back pain with two expert systems. *J. Med. Syst.* **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. *Surg. Neurol.* **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. *J. Electromyogr. Kinesiol.* **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. *Int. J. Neural Syst.* **11**, 335–347 (2001).
49. Abdullah, A. A., Yaakob, A. & Ibrahim, Z. Prediction of spinal abnormalities using machine learning techniques. In *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, 1–6 (IEEE, 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. In *Proc. International Joint Conference on Computational Intelligence*, 455–469 (Springer, 2020).
51. Andrei, D. et al. Computer aided patient evaluation in the low back pain pathology. In *2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics*, 27–30 (IEEE, 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. In *2013 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, 7–12 (IEEE, 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 Netw.* **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. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, 470–475 (IEEE, 2015).
55. Chan, H., Zheng, H., Wang, H., Sterritt, R. & Newell, D. Smart mobile phone based gait assessment of patients with low back pain. In *2013 Ninth International Conference on Natural Computation (ICNC)*, 1062–1066 (IEEE, 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 Conf. Proc.* **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. *J. Med. Syst.* **38**, 50 (2014).
59. Li, H.-X., Wang, Y. & Zhang, G. Probabilistic fuzzy classification for stochastic data. *IEEE Trans. Fuzzy Syst.* **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. *Decis. Support Syst.* **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. *Br. J. Neurosurg.* **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. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 243–249 (IEEE, 2015).
63. Sandag, G. A., Tedry, N. E. & Lolong, S. Classification of lower back pain using K-Nearest Neighbor algorithm. In *2018 Sixth International Conference on Cyber and IT Service Management (CITSM)*, 1–5 (IEEE, 2018).
64. Ung, H. et al. Multivariate classification of structural MRI data detects chronic low back pain. *Cereb. 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. In *International Workshop on Hybrid Neural Systems*, 270–285 (Springer, 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. *Seventh Aust. N.Z. Intell. Inf. Syst. Conf., 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. *Spine J.* **18**, 1540–1551 (2018).
68. Abedi, M. et al. Translation and validation of the Persian version of the STarT Back Screening Tool in patients with nonspecific low back pain. *Man. Ther.* **20**, 850–854 (2015).
69. 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).
70. 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. *Braz. J. Phys. Ther.* **S1413-3555**, 30777–30779 (2019).
71. 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. *Phys. Ther.* **93**, 321–333 (2013).
72. 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. *J. Orthop. Sports Phys. Ther.* **44**, 656–664 (2014).
73. 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. *J. Pain* **16**, 19–30 (2015).
74. 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. *Phys. Ther.* **97**, 561–570 (2017).
75. Bruyere, O. et al. Validity and reliability of the French version of the STarT Back screening tool for patients with low back pain. *Spine* **39**, E123–E128 (2014).
76. 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. *J. Gen. Intern. Med.* **33**, 1324–1336 (2018).
77. Field, J. & Newell, D. Relationship between STarT Back Screening Tool and prognosis for low back pain patients receiving spinal manipulative therapy. *Chiropr. Man. Therapies* **20**, 17 (2012).
78. 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 Emerg. Med.* **25**, 1138–1145 (2018).
79. 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–E936 (2016).
80. George, S. Z. & Beneciuk, J. M. Psychological predictors of recovery from low back pain: a prospective study. *BMC Musculoskelet. Disord.* **16**, 49 (2015).
81. Karran, E. L. et al. The value of prognostic screening for patients with low back pain in secondary care. *J. Pain* **18**, 673–686 (2017).
82. 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 Musculoskelet. Disord.* **16**, 346 (2015).
83. 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 Musculoskelet. Disord.* **20**, 94 (2019).
84. 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. *Spine J.* **19**, 645–654 (2019).
85. 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. *J. Physiother.* **64**, 107–113 (2018).
86. 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. *Man. Ther.* **21**, 120–127 (2016).
87. 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–E979 (2014).
88. 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).
89. Matsudaira, K. et al. The Japanese version of the STarT Back Tool predicts 6-month clinical outcomes of low back pain. *J. Orthop. Sci.* **22**, 224–229 (2017).
90. 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. *J. Orthop. Sports Phys. Ther.* **47**, 314–323 (2017).

91. 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. *Eur. Spine J.* **27**, 2823–2830 (2018).
92. 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? *Eur. J. Pain* **19**, 439–446 (2015).
93. Morso, L. et al. The predictive and external validity of the STarT Back Tool in Danish primary care. *Eur. Spine J.* **22**, 1859–1867 (2013).
94. Morsø, L., Kent, P., Manniche, C. & Albert, H. B. The predictive ability of the STarT Back Screening Tool in a Danish secondary care setting. *Eur. Spine J.* **23**, 120–128 (2014).
95. 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* **41**, 645–652 (2016).
96. 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 Musculoskelet. Disord.* **18**, 345 (2017).
97. 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 Musculoskelet. Disord.* **16**, 201 (2015).
98. Piironen, S. et al. Transcultural adaption and psychometric properties of the STarT Back Screening Tool among Finnish low back pain patients. *Eur. Spine J.* **25**, 287–295 (2016).
99. Pilz, B. et al. The Brazilian version of STarT Back Screening Tool-translation, cross-cultural adaptation and reliability. *Braz. J. Phys. Ther.* **18**, 453–461 (2014).
100. Pilz, B. et al. Construct and discriminant validity of STarT Back Screening Tool—Brazilian version. *Braz. J. Phys. Ther.* **21**, 69–73 (2017).
101. Raimundo, A. M. M. et al. Portuguese translation, cross-cultural adaptation and reliability of the questionnaire "Start Back Screening Tool" (SBST). *Acta. Reumatol. Port.* **42**, 38–46 (2017).
102. 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 Musculoskelet. Disord.* **18**, 360 (2017).
103. 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 Musculoskelet. Disord.* **18**, 232 (2017).
104. 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. *Dan. Med. J.* **65**, A5517 (2018).
105. 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. *Arch. Phys. Med. Rehab.* **99**, 1533–1539 (2018).
106. Tan, C. I. C. et al. Predicting outcomes of acute low back pain patients in emergency department: a prospective observational cohort study. *Medicine* **97**, e11247 (2018).
107. 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. *J. Orthop. Sports Phys. Ther.* **47**, 261–267 (2017).
108. Von Korff, M. et al. Comparison of back pain prognostic risk stratification item sets. *J. Pain* **15**, 81–89 (2014).
109. Yelvar, G. D. Y. et al. Validity and reliablity of Turkish version of STarT Back Screening Tool. *Agri.* **31**, 163–171 (2019).
110. Foster, N. E. et al. Effect of stratified care for low back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. *Ann. Fam. Med.* **12**, 102–111 (2014).
111. Bid, D. D. A study on central sensitization in chronic non specific low back pain. *Indian J. Physiother. Occup. Ther.* **160**, 165–175 (2018).
112. 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. *N. Engl. J. Med.* **339**, 1021–1029 (1998).
113. 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. *Phys. Ther.* **76**, 706–716 (1996).
114. Edmond, S. L. et al. Directional preference, cognitive behavioural interventions, and outcomes among patients with chronic low back pain. *Physiother. Res. Int.* **24**, e1773 (2019).
115. Flavell, C. A., Gordon, S. & Marshman, L. Classification characteristics of a chronic low back pain population using a combined McKenzie and patho-anatomical assessment. *Man. Ther.* **26**, 201–207 (2016).
116. Garcia, A. N. et al. Effectiveness of back school versus McKenzie exercises in patients with chronic nonspecific low back pain: a randomized controlled trial. *Phys. Ther.* **93**, 729–747 (2013).
117. Garcia, A. N., Costa, Ld. 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. *Phys. Ther.* **96**, 623–630 (2016).
118. 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. *J. Orthop. Sports Phys. Ther.* **46**, 514–522 (2016).
119. Johnson, O. E., Adegoke, B. O. & Ogunlade, S. O. Comparison of four physiotherapy regimens in the treatment of long-term mechanical low back pain. *J. Jpn. Phys. Ther. Assoc.* **13**, 9–16 (2010).
120. 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. *Phys. Ther.* **77**, 354–360 (1997).
121. Kilby, J., Stigant, M. & Roberts, A. The reliability of back pain assessment by physiotherapists, using a 'McKenzie algorithm'. *Physiotherapy* **76**, 579–583 (1990).
122. Kilpikoski, S. et al. Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine* **27**, E207–E214 (2002).
123. 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).
124. Long, A., May, S. & Fung, T. The comparative prognostic value of directional preference and centralization: a useful tool for front-line clinicians? *J. Man. Manipulative Ther.* **16**, 248–254 (2008).
125. 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 Med.* **8**, 10 (2010).
126. 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. *J. Man. Manipulative Ther.* **13**, 103–112 (2005).
127. 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. *J. Rehab. Med.* **40**, 858–863 (2008).
128. 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 Musculoskelet. Disord.* **16**, 74 (2015).
129. 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).
130. 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).
131. Razmjou, H., Kramer, J. F. & Yamada, R. Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low back pain. *J. Orthop. Sports Phys. Ther.* **30**, 368–389 (2000).
132. 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).
133. 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. *J. Man. Manipulative Ther.* **10**, 129–135 (2002).
134. Sufka, A. et al. Centralization of low back pain and perceived functional outcome. *J. Orthop. Sports Phys. Ther.* **27**, 205–212 (1998).
135. Werneke, M. & Hart, D. L. Centralization phenomenon as a prognostic factor for chronic low back pain and disability. *Spine* **26**, 758–764 (2001).
136. 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).
137. 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. *J. Orthop. Sports Phys. Ther.* **46**, 726–741 (2016).
138. Werneke, M. W. et al. Directional preference and functional outcomes among subjects classified at high psychosocial risk using STarT. *Physiother. Res. Int.* **23**, e1711 (2018).
139. 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. *J. Man. Manipulative Ther.* **26**, 18–24 (2018).
140. Viera, A. J. & Garrett, J. M. Understanding interobserver agreement: the kappa statistic. *Fam. Med.* **37**, 360–363 (2005).

141. Terwee, C. B. et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J. Clin. Epidemiol.* **60**, 34–42 (2007).
142. Hartvigsen, J. et al. What low back pain is and why we need to pay attention. *Lancet* **6736**, 1–12 (2018).
143. Dolnicar, S. A review of unquestioned standards in using cluster analysis for data-driven market segmentation. In *Conference Proceedings of the Australian and New Zealand Marketing Academy Conference 2002 (ANZMAC), 1–9* (2002).
144. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
145. Fairbank, J. et al. The role of classification of chronic low back pain. *Spine* **36**, S19–S42 (2011).
146. 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 Med. Rev.* **25**, 52–73 (2016).
147. 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. *Disabil. Rehab.* **37**, 1242–1250 (2015).
148. 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. Rehab. Med.* **50**, 143–151 (2014).
149. Steffens, D. et al. Prognosis of chronic low back pain in patients presenting to a private community-based group exercise program. *Eur. Spine J.* **23**, 113–119 (2014).
150. 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).
151. Chou, R. & Shekelle, P. Will this patient develop persistent disabling low back pain? *JAMA* **303**, 1295–1302 (2010).
152. Picavet, H. S. J. Pain catastrophizing and kinesiophobia: predictors of chronic low back pain. *Am. J. Epidemiol.* **156**, 1028–1034 (2002).
153. Ng, S. K. et al. Negative beliefs about low back pain are associated with persistent high intensity low back pain. *Psychol., Health Med.* **22**, 790–799 (2017).
154. Jackson, T., Wang, Y., Wang, Y. & Fan, H. Self-efficacy and chronic pain outcomes: a meta-analytic review. *J. Pain* **15**, 800–814 (2014).
155. 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. *Occup. Environ. Med.* **62**, 851–860 (2005).
156. 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. *J. Orthop. Sports Phys. Ther.* **49**, 698–715 (2019).
157. Kregel, J. et al. Structural and functional brain abnormalities in chronic low back pain: a systematic review. *Semin. Arthritis Rheum.* **45**, 229–237 (2015).
158. Mansour, A. R. et al. Brain white matter structural properties predict transition to chronic pain. *Pain* **154**, 2160–2168 (2013).
159. Van Tulder, M. et al. Chapter 3 European guidelines for the management of acute nonspecific low back pain in primary care. *Eur. Spine J.* **15**, 169–191 (2006).
160. Hayden, J., Dunn, K., Van der Windt, D. & Shaw, W. What is the prognosis of back pain? *Best Pract. Res. Clin. Rheumatol.* **24**, 167–179 (2010).
161. Koes, B. W., van Tulder, M. W. & Thomas, S. Diagnosis and treatment of low back pain. *BMJ* **332**, 1430–1434 (2006).
162. Alrwaily, M. et al. Treatment-based classification system for low back pain: revision and update. *Phys. Ther.* **96**, 1057–1066 (2016).
163. Lohr, K. N. Assessing health status and quality-of-life instruments: Atributes and review criteria. *Qual. Life Res.* **11**, 193–205 (2002).
164. Andresen, E. M. Criteria for assessing the tools of disability outcomes research. *Arch. Phys. Med. Rehab.* **81**, S15–S20 (2000).
165. Wells, G. et al. The Newcastle−Ottawa Scale (NOS) for assessing the quality if nonrandomized studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (2016).

ACKNOWLEDGEMENTS

S.D.T. is supported by an Australian Government Research Training Program (RTP) Scholarship.

AUTHOR CONTRIBUTIONS

S.D.T.: Conception of review, data extraction (AI/ML, McKenzie and STarT Back), risk of bias assessment, preparation and revision of manuscript. M.A.: Feedback, guidance and revision of manuscript. X.Z.: Data extraction (McKenzie), revision of manuscript. P.J.O.: Feedback, guidance and revision of manuscript. C.T.M.: Feedback, guidance and revision of manuscript. T.W.: Feedback, guidance and revision of manuscript. D.L.B.: Conception of review, database searches, data extraction (AI/ML and STarT Back), risk of bias assessment and revision of manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41746-020-0303-x.

Correspondence and requests for materials should be addressed to S. D.T. or D. L.B.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.