ELSEVIER

The SPINE JOURNAL

The Spine Journal 000 (2024) 1−14

Systematic Review / Meta-Analysis

Machine learning in the diagnosis, management, and care of patients with low back pain: a scoping review of the literature and future directions

Andreas Seas[a,b], Tanner J. Zachem[a,c], Bruno Valan[d,e], Christine Goertz[e], Shiva Nischal[f], Sully F. Chen[a], David Sykes[a], Troy Q. Tabarestani[a], Benjamin D. Wissel[a], Elizabeth R. Blackwood[g], Christopher Holland[h], Oren Gottfried[a,e], Christopher I. Shaffrey[a,e], Muhammad M. Abd-El-Barr[a,e,*]

[a] *Department of Neurosurgery, Duke University Medical Center, Durham, NC, USA*
[b] *Department of Biomedical Engineering, Duke Pratt School of Engineering, Duke University, Durham, NC, USA*
[c] *Department of Mechanical Engineering, Duke Pratt School of Engineering, Duke University, Durham, NC, USA*
[d] *Duke University Medical Center, Duke Institute for Health Innovation, Durham, NC, USA*
[e] *Department of Orthopaedic Surgery, Duke University Medical Center, Durham, NC, USA*
[f] *Department of Neurosurgery, University of Cambridge School of Clinical Medicine, Cambridge, England, UK*
[g] *Duke University Medical Center Library & Archives, Durham, NC, USA*
[h] *Carolina Neurosurgery & Spine Associates, Charlotte, NC, USA*

Received 2 May 2024; revised 19 August 2024; accepted 14 September 2024

Abstract

BACKGROUND CONTEXT: Low back pain (LBP) remains the leading cause of disability globally. In recent years, machine learning (ML) has emerged as a potentially useful tool to aid the diagnosis, management, and prognostication of LBP.

PURPOSE: In this review, we assess the scope of ML applications in the LBP literature and outline gaps and opportunities.

STUDY DESIGN/SETTING: A scoping review was performed in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines.

METHODS: Articles were extracted from the Web of Science, Scopus, PubMed, and IEEE Xplore databases. Title/abstract and full-text screening was performed by two reviewers. Data on model type, model inputs, predicted outcomes, and ML methods were collected.

RESULTS: In total, 223 unique studies published between 1988 and 2023 were identified, with just over 50% focused on low-back-pain detection. Neural networks were used in 106 of these articles. Common inputs included patient history, demographics, and lab values (67% total). Articles published after 2010 were also likely to incorporate imaging data into their models (41.7% of articles). Of the 212 supervised learning articles identified, 168 (79.4%) mentioned use of a training or testing dataset, 116 (54.7%) utilized cross-validation, and 46 (21.7%) implemented

FDA device/drug status: Not applicable.

Author disclosures: ***AS:*** Nothing to disclose. ***TJZ:*** Nothing to disclose. ***BV:*** Nothing to disclose. ***CG:*** Consulting: Whole Health Institute (E), George Family Foundation (B); Speaking and/or Teaching Arrangements: Georgetown University (A), Edward Via College of Oesteopathic Medicine (B), European Chiropractic Union (B), Palmer College of Chiropractic (B), Canadian Chiropractic Association (B), AAOMPT (B), Northwestern Health Sciences (B), University Hospitals (B); Trips/Travel: AHCJ (A); Board of Directors: Logan University (B); Research Support (Investigator Salary, Staff/Materials)^: NIH (40% salary support); Grants: NIH (H, Paid directly to institution/employer). ***SN:*** Nothing to disclose. ***SFC:*** Nothing to disclose. ***DS:*** Nothing to disclose. ***TQT:*** Nothing to disclose. ***BDW:*** Nothing to disclose. ***ERB:*** Nothing to disclose. ***CH:*** Royalties: Globus Medical (None); Consulting: Globus Medical (None), Medtronic (None), SpineWave (None), Instrinsic Therapeutics (None); Speaking and/or Teaching Arrangements: Medtronic (None), Globus Medical (None); Research Support (Investigator Salary, Staff/Materials)^: Medtronic (None). ***OG:*** Nothing to disclose. ***CIS:*** Nothing to disclose. ***MMAEB:*** Nothing to disclose.

*Corresponding author. Duke Neurosurgery, Duke University Medical Center, 40 Duke Medicine Circle, Ste 1554, Durham, NC 27710, USA.

E-mail address: muhammad.abd.el.barr@duke.edu (M.M. Abd-El-Barr).

https://doi.org/10.1016/j.spinee.2024.09.010

hyperparameter optimization. Of all articles, only 8 included external validation and 9 had publicly available code.

CONCLUSIONS: Despite the rapid application of ML in LBP research, a majority of articles do not follow standard ML best practices. Furthermore, over 95% of articles cannot be reproduced or authenticated due to lack of code availability. Increased collaboration and code sharing are needed to support future growth and implementation of ML in the care of patients with LBP.

Keywords: Data science; Integrative medicine; Low back pain; Machine learning; Artificial intelligence; Neural networks

Introduction

Low back pain (LBP) remains the leading cause of years lived with disability (YLDs) globally [1]. Despite advances in diagnostic and therapeutic technology, the burden of LBP continues to grow at an alarming rate, with a 60% increase in cases from 1990 to 2020 [1,2]. The global financial impact of LBP is estimated between $50 and $100 billion USD per annum. In 2016 alone, low back pain and neck pain were identified as the largest direct healthcare costs in the US [3]. The combined high cost and widespread burden of disability make efficient diagnosis and management of LBP a priority for patients, health systems and practitioners worldwide [4,5].

The presentation of LBP can vary significantly between patients [6]. There are also many diverse treatment modalities for LBP, ranging from conservative nonsurgical interventions such as physical therapy and chiropractic management, to pharmacologic-based therapies, to physiatric or surgical interventions [7−10]. This complexity makes identifying and implementing the right treatment regimen exceedingly challenging [11]. Additional compounding factors affecting patient care include socioeconomic and psychosocial variables, as well as the relatively low rate of guideline concordant care within LBP [12−14].

The complexity of LBP care is exacerbated by several important factors. First, the traditional biomedical approach to managing LBP does not align well with the more comprehensive biopsychosocial model of pain [15−17]. In the latter more updated approach, patient pain is considered in the context of factors beyond biomechanics and neurologic change, such as patient social and cultural background, and psychologic state [18]. Second, there is a significant gap between common treatment approaches and guideline-concordant care [15,19]. Recent guidelines from the American College of Physicians (ACP) recommend that patients with LBP progress systematically through therapies from least-to most-invasive [15]. By this logic, a vast minority of patients would require prescription of opioids, with surgical intervention reserved for those patients with severe, unremitting pain recalcitrant to conservative therapy and those who exhibit red flag signs: incontinence, anesthesia in the groin area, and rapidly progressing neurologic deficits. Despite this, inappropriate outpatient referrals and inpatient consultations to surgical services are abundant, with over 80% of referrals deemed inappropriate in a recent study [20]. This mismanagement of patient care poses a significant financial burden to health systems globally [21−25]. One of the potential contributors to this guideline-discordance is lack of awareness of these guidelines on the part of patients and physicians [26].

Machine learning (ML) has emerged in recent years as a tool to assist in answering complex questions. As such it has seen application in the diagnosis, management, and prognostication of a variety of musculoskeletal and neurosurgical pathologies [27−30]. ML algorithms can capture complex, nonlinear relationships in data, making them ideal to help in challenging clinical scenarios such as LBP care. However, these same characteristics make ML tools challenging to develop and integrate into every-day use. The objective of this study is to evaluate the state of ML applications in the care of patients with LBP and propose future directions for ML-guided integrative LBP care.

Methods

Article identification

A scoping review was performed to clarify the breadth of research publications utilizing machine learning in the evaluation and/or clinical management of low back pain. This was performed in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist, which can be found in Appendix A [31]. The review was registered on Open Science Framework (doi **redacted for anonymity**). A comprehensive search strategy was created with our research librarian team members (EB) to include terms related to low back pain and machine learning (Appendix B). These terms were searched on the Pubmed, Embase, Web of Science, and IEE XPlore databases on April 1st of 2023, and metadata of all resulting articles until that date were retrieved. Only English language articles were reviewed. No date restrictions were applied. These articles were imported into Covidence Extraction 2.0 systematic review software (Veritas Health Innovation, Melbourne, Australia). Duplicate articles were removed through automated filtration within Covidence followed by manual identification. In total, 2282 non-duplicate citations were screened.

Article screening

Screening was performed in two stages: title/abstract screening followed by full text screening. In both stages, reviewers assessed article adherence to pre-specified inclusion and exclusion criteria, as given in Appendix C. Screening was performed independently by two members of the review team (AS, TZ). Conflicts between reviewers were subsequently resolved by a third independent team member (DS). Inter-rater reliability was assessed via calculation of a Cohen's Kappa score for both the title/abstract screening stage and the full-text screening stage. A standard error was also computed at each stage and used to compute a 95% confidence interval for each Cohen's Kappa score.

Data extraction, analysis, and processing

Two independent reviewers collected several data elements from each publication. Articles and their corresponding algorithms were first placed into one or more "stages" of LBP management or intervention. Articles categorized as "Risk Prediction" aimed to predict future incidence of LBP in patients without a current diagnosis of LBP. The "Detection" category included algorithms aimed at detecting LBP in patients with an existing diagnosis, perhaps through imaging or clinical characteristics. Algorithms classified as "Characterization" focused on subclassifying or stratifying LBP, and included tools meant to further our understanding its pathophysiology. "Triage" algorithms used imaging or other data to predict patient nonsurgical referrals/appointments, while "Intervention" algorithms predicted surgical treatment. "Outcome Prediction" algorithms were those built to predict patient outcomes from a specific intervention whether that be surgical or non-surgical. These include but are not limited to post-treatment mortality, LBP severity, readmission, and pain recurrence.

Each included article built or tested at least one ML algorithm, broadly defined here as any mathematical formulation that uses past data to predict a future event (with the exception of rules-based systems) [32]. Linear regression models assume a linear relationship between a set of input predictors and a continuous output. An example of a continuous output to predict for a LBP patient could be length of stay after surgery or LBP severity [33]. Logistic regression models are like linear regression models except their output is the probability of a binary outcome, such as whether a patient will be readmitted within 30 days of a lumbar fusion or have a postoperative complication [34]. Decision tree models are flowchart-like models where each node represents a feature, each branch represents a decision, and each leaf represents an outcome [35]. Random forests are built by creating several decision trees from the same data and combining their predictions [36]. Extreme Gradient Boosting (XGBoost) algorithms involve building multiple decision trees one after the other, where each tree "corrects" the errors of its "ancestors" rather than averaging across all trees like in the random forest model [37]. Support vector machines (SVMs) take input data and find the best boundary (also called a "hyperplane") that separates one group of data from another [38]. Naïve-Bayes models use the probability of each model input being associated with an outcome to make predictions with all the model inputs together [39]. Clustering models group similar datapoints together in an unsupervised approach, without knowledge of whether these groupings are relevant or meaningful [40]. They can be used to explore data and even segment radiologic images. K-Nearest Neighbors models make predictions from model inputs based off results from neighboring datapoints, here meaning similar scans or similar patients [41]. Natural Language Processing (NLP) is a larger subfield of the ML world referring to the use of algorithms to process and analyze raw text [42]. In this review, we specifically classified articles as using NLP if they used raw text to make predictions relevant to LBP care, and we did not include algorithms that used NLP to simply extract data. Finally, neural networks are some of the most common ML algorithms and are modeled after the human brain. They consist of layers of interconnected "neurons" connected through linear or nonlinear "axons" with adjustable weights [43].

All the ML algorithms noted above require input of data. Various types of data inputs were recorded from each article type, mainly grouped into imaging and nonimaging inputs. Imaging data inputs included Magnetic Resonance Imaging (MRI) scans, Computed Tomography (CT) scans, X-rays, functional MRI (fMRI), or other imaging inputs. We also noted if raw imaging report text was used as a model input, or if specific data were manually extracted from images, such as spinopelvic parameters. Nonimaging data included patient demographics, lab values and history, as well as posture and balance data or kinematic data from movement.

Model outputs were also recorded. These included binary data such as risk of developing LBP in the future, present LBP, LBP recurrence, imaging abnormality, complications after intervention, need for surgery, need for physical therapy, recommendation of a specific therapy, readmission, reoperation, and mortality. Continuous outputs included patient reported outcome measures (PROMs), LBP severity, patient satisfaction, and length of stay. Imaging segmentation was the only recorded output that is neither binary nor continuous.

Various model characteristics and best practices were recorded for each algorithm [44]. We recorded first whether models were supervised or unsupervised: supervised models are built using data with labeled outputs (like the y-value in a linear regression), while unsupervised models are trained without guidance on what the output should be. It was also noted whether data were split into a training and a testing set which don't share data, a common practice in ML model development. Models are built using the training dataset, and their performance is assessed on the testing data. This prevents model overfitting, a phenomenon where a model learns to predict noise rather than relevant signal, and doesn't generalize well to new, unseen data. Cross validation is another common method to prevent overfitting which was assessed in our extraction. It involves training and testing a model iteratively on different subsets of the same dataset to see how well it generalizes to unseen data [45]. We also assessed how many models implemented hyperparameter optimization techniques, where different model architectures are tested to identify which one(s) work best to predict a specific output [46]. An example of this would include training several different neural networks to predict hospital readmission on a dataset, each network with a different number of neurons, and picking the best network from the bunch. After collecting all these data, we assessed whether any of these models were externally validated on a unique set of data outside of a train-test split, such as data from a different hospital. We finally identified whether code was made available for any of these articles on Github or other similar code-sharing platforms.

Once raw article data were collected and cleaned, they were analyzed within Python 3.9.18 using numpy 1.22.3, pandas 1.4.2, matplotlib 3.5.1, and seaborn 0.12.2 [47−50].

Results

Of the 2,282 unique studies extracted, 346 met inclusion and exclusion criteria on title/abstract screening (Cohen's Kappa = 0.706, 95% confidence interval from 0.666 to 0.746). Of these, 123 were removed in full-text review due to violation of inclusion and exclusion criteria, leaving 223 studies as illustrated in Fig. 1 (Cohen's Kappa=0.337, 95% confidence interval from 0.239 to 0.435).

Applications of ML in LBP

The studies included in this review described applications of machine learning at every stage of low back pain management from predicting a patient's risk of developing LBP, to predicting the outcome of intervention for chronic LBP (Fig. 2A). A majority of studies focused on the detection of LBP (52%), followed by outcome prediction (24.7%) and disease characterization (8.1%). Only six articles attempted to triage patients to different types of therapeutic approaches for their pain.

There are two periods of rapid expansion in this research topic, and the time periods of articles included 1988-2003, and 2010 to the present day, with a plateau from 2003 to 2010 (Fig. 2B, C). Algorithms aimed at LBP detection were the first to emerge in the late 1990s and into the 2000s, making up a majority of articles from 1988-2003. However, the second period of expansion from 2010 to present day has included algorithms with several unique applications in LBP such as predicting proper patient triage or intervention, or even future development or worsening of LBP. However, a large portion of these algorithms have aimed at predicting outcomes of LBP treatment such as patient mortality, pain recurrence, medication use, or hospital admission and readmission.

ML algorithms in LBP

Neural networks are the most commonly reported ML method among the included articles and were utilized in 47.5% of all articles (Fig. 3A, B). They were also the first to emerge at the beginning of the 1990s. They have since continued to grow in use but have been followed by other simpler algorithms including logistic regression (24.7%), support vector machine (25.1%), and random forest (22.9%), all of which have grown in use since 2015.

Common model inputs and outputs

All early studies used nonimaging data as model inputs, including pertinent patient history, demographics, and laboratory values (Fig. 4C). However, imaging data began to appear in models in the early 2010s, with MRI leading all other imaging data inputs (Fig. 4A). At present, 41.7% of all studies include some imaging data as an input. In recent years, other types of non-imaging data have made their way into the literature including kinematic and posture data.

The most common output predicted in the included studies was the presence of low back pain itself (29.6%, Table 1). This was followed by the prediction of imaging abnormality (26.0%), and the possibility for recurring LBP after treatment (15.2%).

Machine learning types and best-practices

The overwhelming majority of studies, (96.0%) utilized supervised learning, with only 17 total studies reporting use of unsupervised learning techniques (Fig. 5). Of the supervised learning articles, 168 (79%) clearly implemented a split training/testing protocol, with 116 (52%) utilizing cross-validation to minimize the risk of overfitting. A total of 46 studies implemented hyperparameter optimization techniques to ensure peak model performance. Thirty articles reported model discrimination data, with 16 reporting on model calibration. Very few models shared in these studies were also publicly available via Github or other knowledge-sharing platform (9/223), and only eight models were externally validated.

Trends in model use across stage, inputs, outputs, and practices

Various trends are observed when cross-tabulating ML model type with various article features. As seen in Fig. 6A, detection of LBP was most performed using neural networks, whereas outcome prediction was most frequently performed using logistic regression models. Most imaging analysis was performed with neural networks, especially with MRI data (Fig. 6B). However, manually extracted data was more evenly distributed across ML methodologies. The same is true for nonimaging data which was an input for a

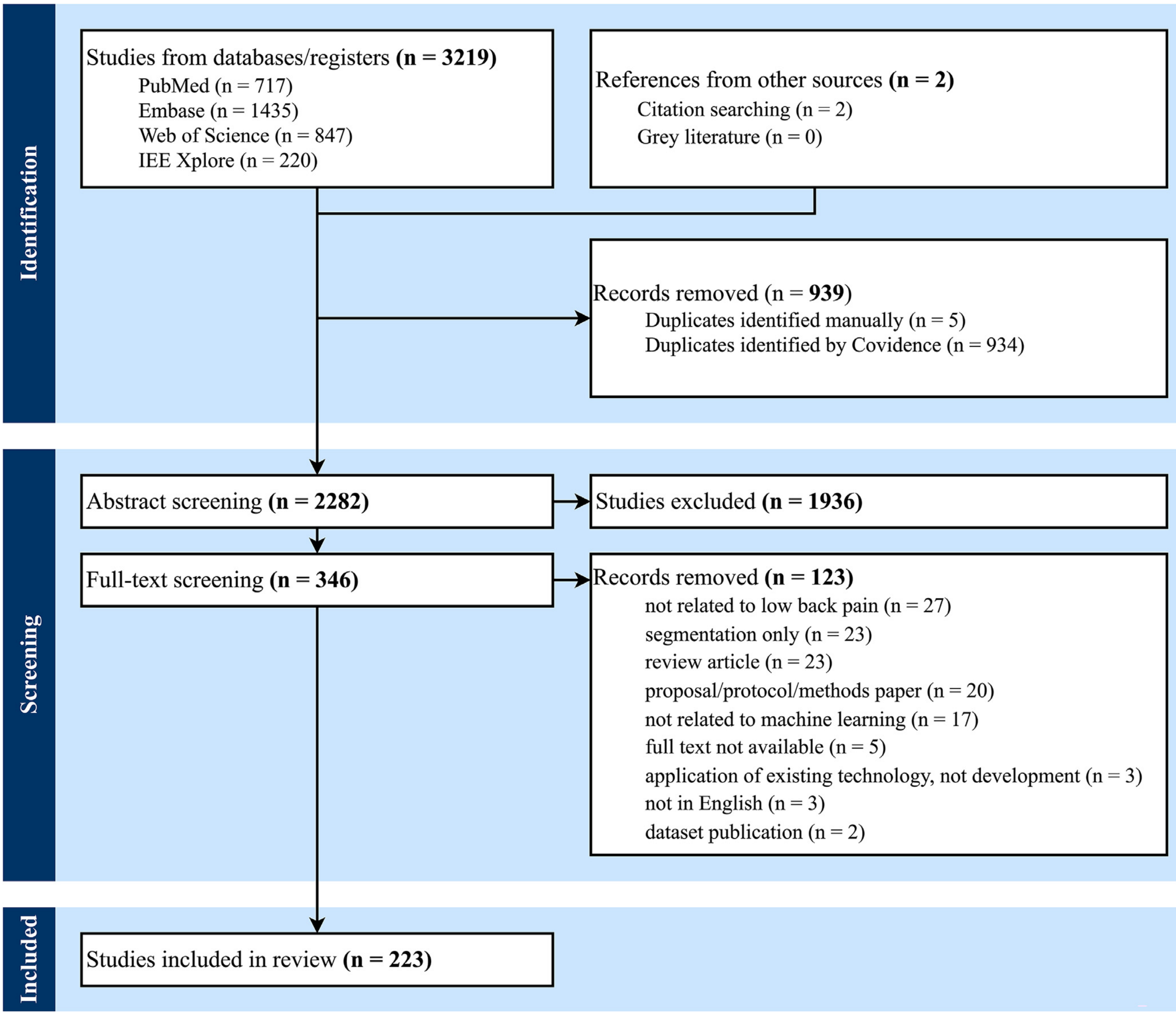

Fig. 1. Consort diagram illustrating the process for article collection, screening, and final inclusion.

wide range of models (Fig. 6C). Outputs were dispersed across all model types, and ML best practices were distributed across all models (Fig. 6D−E).

Discussion

In this scoping review, we share data on the stage, inputs, outputs, methods, and availability of machine learning tools built for the diagnosis, management, and care of patients with LBP.

Machine learning is growing exponentially in the low back pain literature

The use of ML in the LBP literature is rising rapidly. This rise can be subdivided into three domains: slow growth from 1988 to 2003, a plateau from 2003 to 2010, and exponential expansion from 2010 to the present day. This trend can be explained by the recent democratization of machine learning and high-performance computing, the expanding use of electronic health records, and the rising popularity of machine learning within the public domain as well as in the research of academics across the medical space.

The first papers in this review were released in the 1980s and 90s, included fewer than 400 datapoints, and all but one article utilized neural networks [51−58]. At that time, all data were manually tabulated from written records and computing resources were incredibly limited and inflexible. The challenges inherent in data collection and model development at that time also limited model utility: a tool requiring a custom computer would not have been practical, nor would it have been feasible in the clinical setting regardless of its predictive ability.

Since the 1990s, several free and open tools have emerged allowing the implementation of ML by nearly

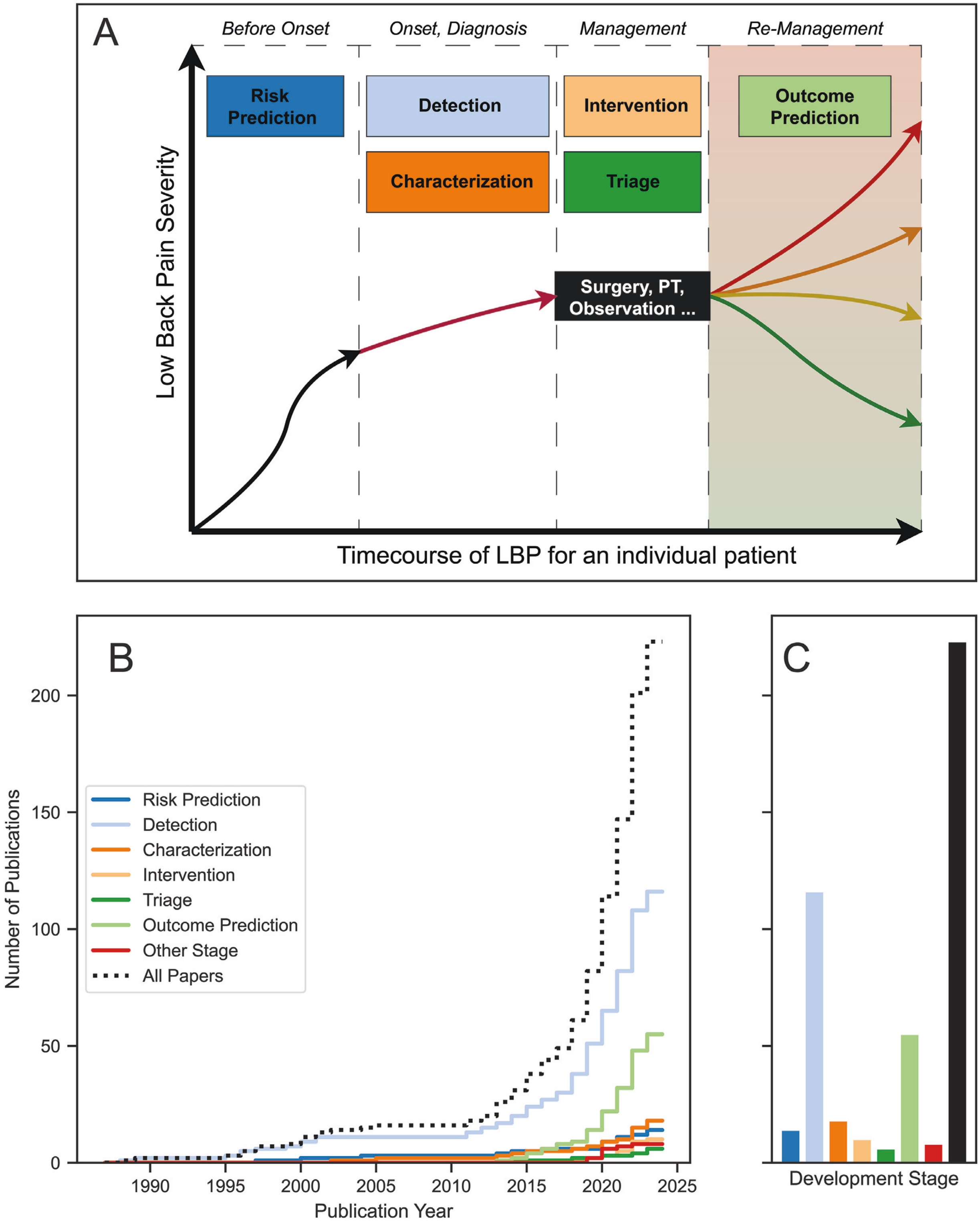

Fig. 2. Breakdown of publications and how their algorithms fit into a general framework of low back pain machine learning and treatment. (A) This is a conceptualization of the path of an individual patient across time, and the various points where a machine learning algorithm could be applied (represented by the various colored boxes). Patient time course is broken into four phases, "Before Onset" when a patient still does not have constant LBP but may have risk factors, "Onset, Diagnosis" when a patient first has constant LBP but is not yet receiving treatment, "Management" when a patient first sees a provider and undergoes some treatment whether that be pharmacologic, physiotherapeutic, or surgical, and "Re-Management" when a patient is continuing management whether that is escalation or de-escalation of care. Each arrow within the "Re-Management" phase represents a different path a patient's pain could take, from significantly improving (green arrow) to significantly worsening (redarrow). (B) The breakdown of articles and their respective algorithms based on their stage of intervention, and (C) the cumulative sum of these publications over time.

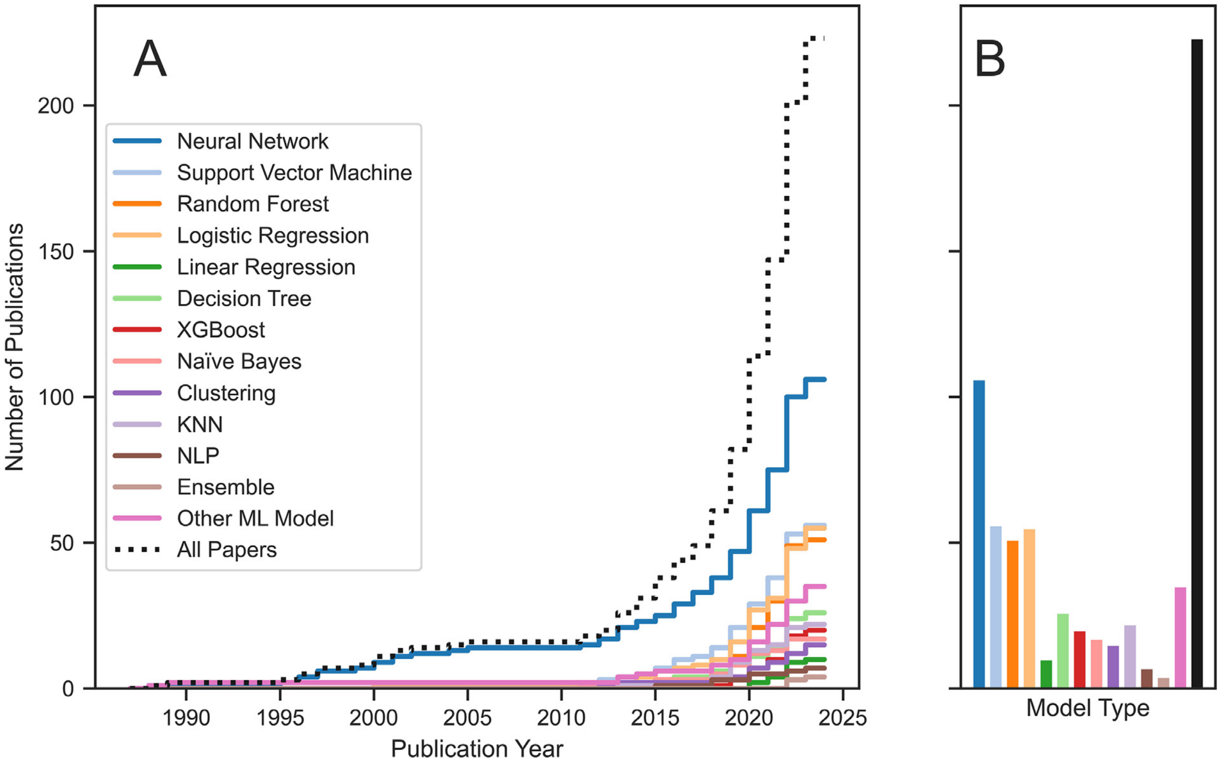

Fig. 3. Breakdown of (A) the algorithms employed and their use over time, and (B) cumulatively.

every individual with access to the internet and a modern computer. Scikit-learn, one of the predominant Python packages for machine learning, was first released in 2010 and provides easy implementation of hundreds of algorithms including but not limited to neural networks [59]. The release of this package and other open-source tools (which have freely available and modifiable source code) may explain the rapid rise in publications seen around 2012 in Figs. 2B and 3A, as well as the emergence of non-neural-network algorithms such as support vector machines and random forest classifiers. In the later 2010s, other open-source tools emerged allowing for more complex analysis, including Tensorflow and Pytorch. Both packages are intended for deep learning, a critical tool for image processing involving the use of multi-layered (or "deep") neural networks [60,61]. The slightly delayed release of these tools tracks with the similarly delayed growth of models using imaging inputs, as shown in Fig. 4A. In the past few years, several groups have worked to automate machine learning entirely, with specific applications in improving the efficiency of diagnostic radiology and disease prediction in the backend of electronic health record systems [62,63]. Several large companies have also built interfaces for no-code machine learning including Amazon with SageMaker Canvas and Microsoft with Azure [64,65]. While still early in their use, these tools have seen application in varied scenarios including automated aphasia assessment and modeling of drug mechanism of action [66,67].

In addition to improved access to ML tools, recent growth in ML in LBP is also likely a result of expanding use of electronic health record (EHR) systems, especially at academic and university hospitals. The Medicare and Medicaid EHR Incentive Programs was introduced in 2011 to increase the utilization of EHR systems across the country [68,69]. As a result, the percent of US hospitals utilizing EHR systems jumped from 59% in 2013 to 97% in 2014, and has remained at 96% since then [70]. This may have accelerated the expansion of ML literature in LBP seen beginning in 2015. The next step in this expansion is the integration of ML into electronic health records systems. Recent work by Jiang and Oermann has shown the immense power of health system-scale machine learning. Specifically, they trained and tested language models on unstructured clinical note data from their EHR and were able to successfully predict various important patient outcomes including in-hospital mortality, 30-day readmission, and even insurance denial [71].

Another key reason for the exponential growth of ML in LBP since 2010 is the increased popularity of ML in the scientific community. According to a review by Pugliese et al. [72] the use of ML in publications has grown exponentially across nearly every domain from medicine to law to cybersecurity. Several journals have even created new sub-journals or divisions for machine learning, including New England Journal of Medicine - Artificial Intelligence, Nature Machine Intelligence, and more [73]. This

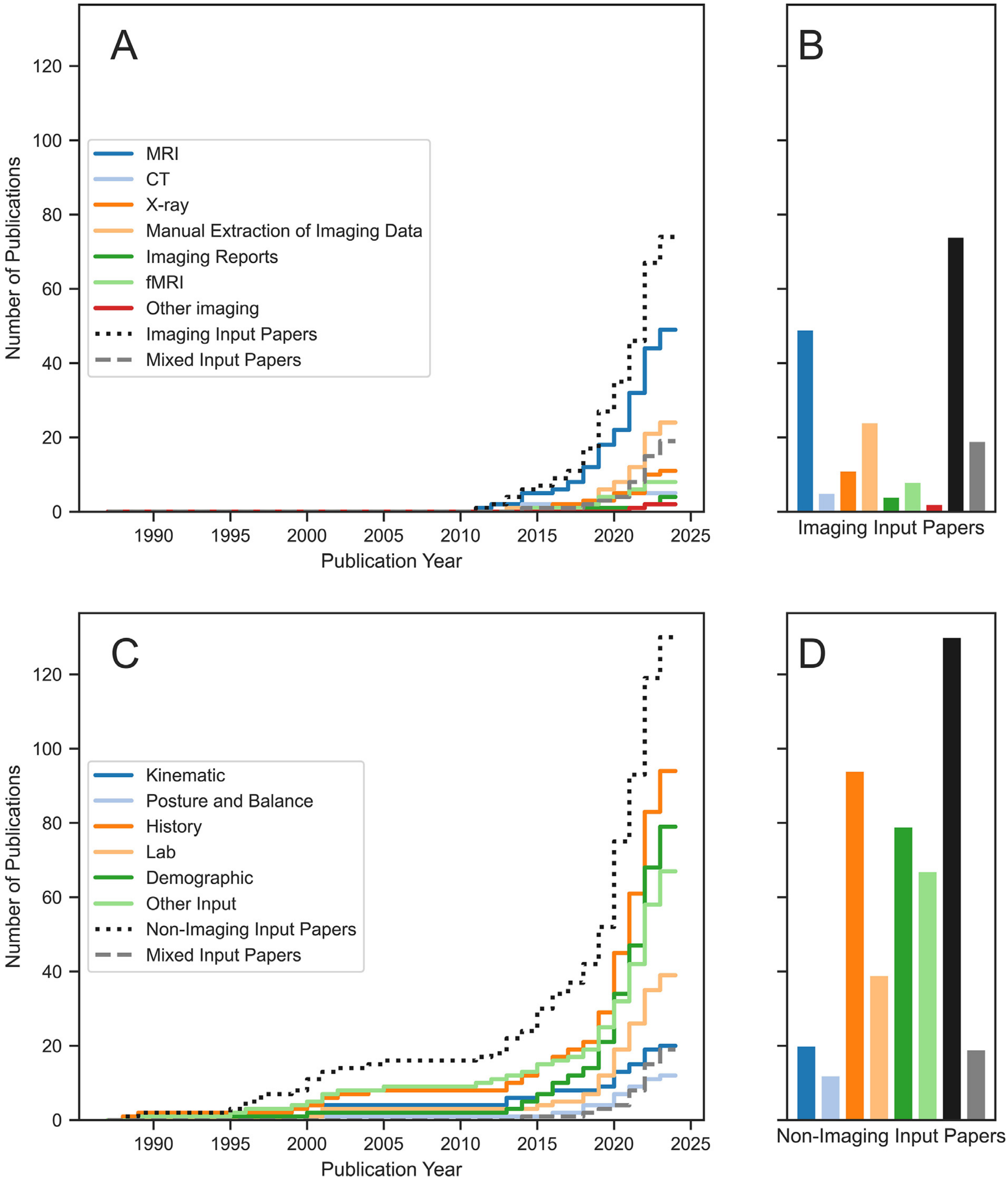

Fig. 4. Illustration of (A) yearly and (B) cumulative counts of publications using imaging inputs, as well as (C) yearly and (D) cumulative counts of publications using other types of inputs including patient history, lab values, etc.

simultaneous expansion in supply of and demand for ML literature is a likely precursor to continued exponential growth [74]. It also underscores the need for expansion in regulation and review of these ML applications, as well as the development of systems for clinical integration and prospective validation [75−78].

We need more work in integrative low back pain machine learning and triage

Many of the articles and models included in this review assess a single element or outcome in a patient's LBP management path (Fig. 1A). Examples include models that diagnose LBP from MRI images or predict postoperative

Table 1
Most common model outputs amongst all articles

Model outputs	
LBP* Presence	66 (29.6%)
Imaging abnormality	58 (26.0%)
LBP recurrence	34 (15.2%)
PROMs2†	19 (8.5%)
General Complications	18 (8.1%)
LBP Severity	17 (7.6%)
Risk of developing LBP	14 (6.3%)
Need for Surgery	13 (5.8%)
LBP Diagnosis/Cause	13 (5.8%)
Patient satisfaction	10 (4.5%)
Therapy to use	7 (3.1%)
Segmentation	6 (2.7%)
Readmission	6 (2.7%)
Length of Stay	5 (2.2%)
Reoperation	2 (0.9%)
Need for physical therapy	1 (0.4%)
Patient Mortality	1 (0.4%)
Other	6 (2.7%)
Total # Articles	223

* Low Back Pain.
† Patient Reported Outcome Measures.

outcomes based on patient weight [79,80]. However, only six models (2.7%) were built to aid in patient triage or integrative care across medical and surgical specialties (Fig. 1C) [81−86]. Furthermore, a majority of outcome-prediction models studied outcomes only after surgical interventions. The few articles which explored patient outcomes after rehabilitation-based interventions predicted outcomes such as pain reduction or change in patient reported outcomes after epidural injections, consistent physiotherapy, and other non-surgical interventions [87−89]. The management of LBP is becoming more complex and more integrative, with spine centers hiring surgeons, pain physicians, physical therapists, and chiropractors to care for various patient needs [26,90]. This complexity, combined with the aforementioned paucity of research on rehabilitation-based interventions outlines the need for more ML research to understand the wider spectrum of care, developing tools to aid in triage and identification of effective non-surgical treatment modalities, as done by Nijeweme-d'Hollosy, Purohit, Knoop, and others in the literature [85−87].

There is a need for code sharing, external validation and prospective implementation

While machine learning has exploded in the LBP literature, many models don't implement machine learning best-practices. Approximately one of every five articles implementing supervised learning in this study did not report splitting their data into a training and testing set. This is a critical step in the development of any predictive model, as it allows for unbiased assessment of performance against data not used in its development [91]. Nearly half of the supervised learning models did not include cross-validation, another important step in model development chiefly aimed to prevent overfitting to the test dataset [91]. Only 46 of all included articles mentioned hyperparameter optimization, a practice which ensures that models provide the best possible prediction.

Of all 223 articles, only nine provided the code used to generate their results. This paucity of open code sharing is unfortunately unsurprising: Hamilton et al. [92] report both the declared and actual availability of data and code in the medical literature at less than 0.5%. This low availability of data holds true even in the radiology artificial intelligence literature, an area of medicine that sports multiple links with the computer science realm through the Brain Tumor Segmentation Challenge (BraTS) and the Abdominal Trauma Detection AI

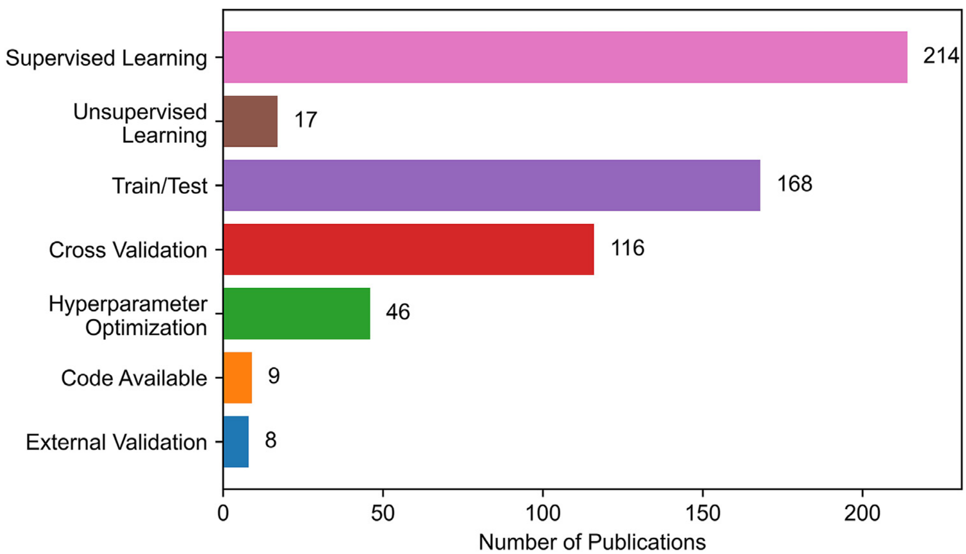

Fig. 5. Breakdown of the prevalence of ML learning types and best practices across all 223 manuscripts.

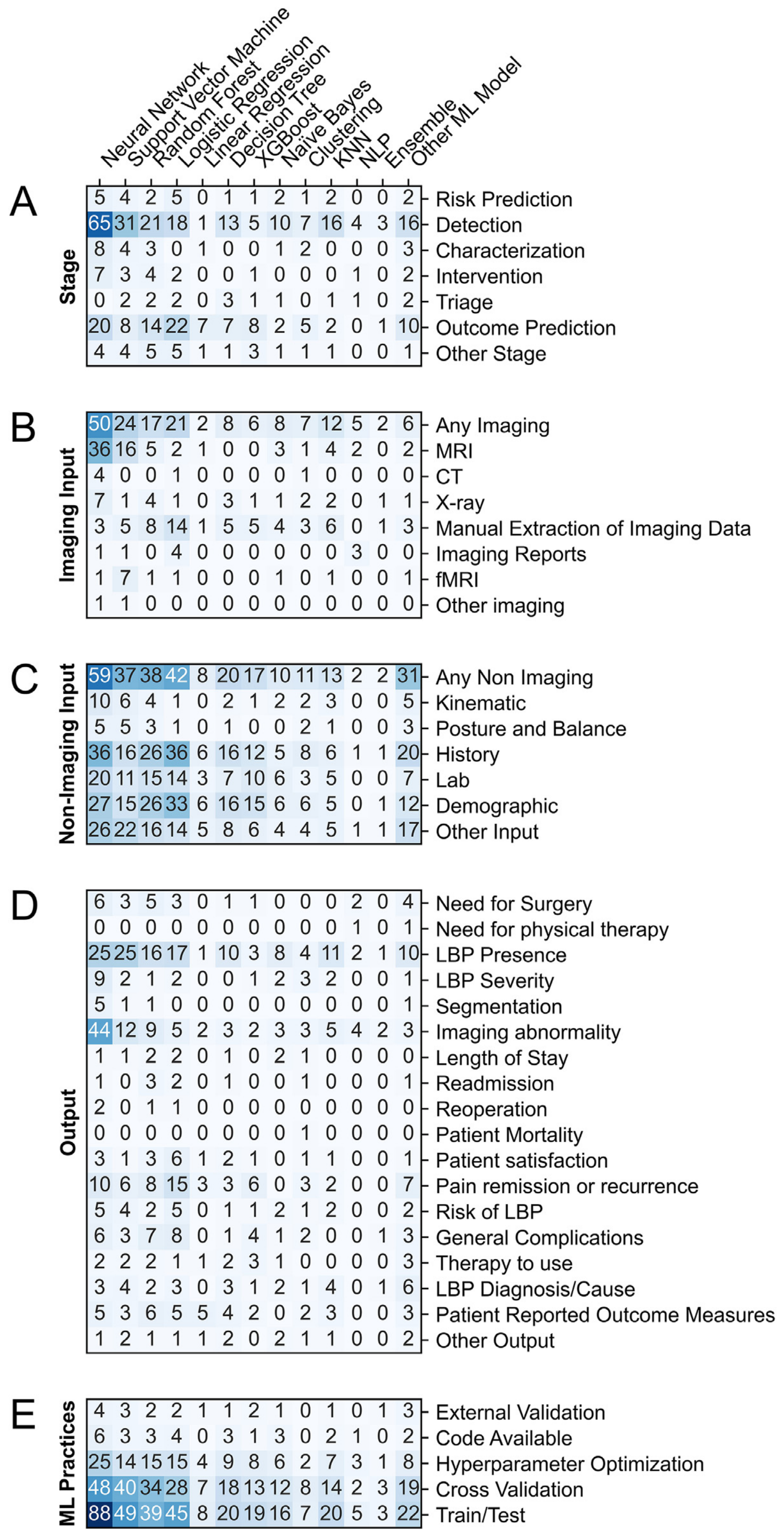

Fig. 6. Cross-tabulated heatmaps of articles showing (A) article stage, (B) imaging input, (C) Nonimaging Input, (D) Output, and (E) ML Practices against the type of model used. Darker blue shade corresponds to a higher number of articles using that model for a particular stage, input, etc.

Challenge to name a few [93−95]. The low availability of code is likely related to the corresponding low number of studies with external validation found in the LBP ML literature (8 total). It is important to note that privacy concerns may play a role in the lack of data availability and code sharing, both with regards to ensuring patient confidentiality and protecting the intellectual property of industry partners.

The observed variability in ML practices and lack of code availability point underscore the need for scientists, clinicians, and publishers alike to adopt guidelines for reporting machine learning research in LBP and other clinical areas. To this end, several reporting guidelines have been established for development and validation of ML tools, and clinical trials involving ML [76,77,96].

Future directions

Despite the proliferation of ML based research aiding the diagnosis and treatment of low back pain, few models are ever deployed into clinical practice. This conspicuous gap in development and implementation is not isolated to lower back pain but prevalent across the spectrum of AI in healthcare. The transition from machine learning research to clinical integration, is fraught with challenges beyond model development, including the need for specialized expertise in data infrastructure, software engineering, and ethical considerations that become exceedingly important when ML influences real-time clinical decisions.

Limitations

This study has several limitations, including the use of modern terminology for machine learning. As with any fast-paced field, the terminology surrounding ML changes rapidly, and the names of several tools have changed with time. For example, what we now call "deep neural networks" were originally called "multilayer perceptrons", and this shifting terminology may have contributed to articles not being included in review. However, most modern search engines including those searched in this review include key terms and labels which are constantly updated to include both past and present names of similar/identical search terms. Another limitation in this work is the fact that data are only included to April 1st, 2023. Given the exponential growth of this field, a significant number of additional articles may have been added since the initial search was performed. We did not extract data regarding whether ML models were trained on socioeconomic or psychosocial factors, and this is an area for future exploration. Finally, we assume that if ML best-practices were used in the development of models, they were reported within the text of each article. However, this is an imperfect assumption, as some practices including cross validation, hyperparameter optimization, and calibration are either automatically implemented, or not reported in the clinical literature.

Conclusion

Machine learning is a continually growing element of the low back pain literature and medicine at large and has expanded with the democratization of ML techniques. Its applications to date have focused on detection of LBP and prediction of outcomes of surgical interventions, but few models have been built to assess the impact of nonsurgical or integrative care for LBP. Few studies utilize standard ML best-practices including train/test splitting, cross-validation, and hyperparameter optimization. Fewer than ten studies to date have shared their code and methods openly with the scientific community. Future work in the application of ML to LBP care must expand to include elements across the spectrum of guideline-concordant care and be shared openly to promote external validation and efficient implementation in clinical practice.

Declaration of competing interest

One or more of the authors declare financial or professional relationships on ICMJE-TSJ disclosure forms.

CRediT authorship contribution statement

Andreas Seas: Writing − review & editing, Writing − original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tanner J. Zachem:** Writing − original draft, Data curation, Conceptualization. **Bruno Valan:** Writing − review & editing, Writing − original draft, Methodology, Data curation, Conceptualization. **Christine Goertz:** Writing − review & editing, Writing − original draft, Supervision, Project administration, Methodology, Data curation, Conceptualization. **Shiva Nischal:** Writing − original draft, Data curation. **Sully F. Chen:** Writing − review & editing, Software, Data curation. **David Sykes:** Writing − review & editing, Data curation, Conceptualization. **Troy Q. Tabarestani:** Writing − review & editing, Data curation, Conceptualization. **Benjamin D. Wissel:** Writing − review & editing, Writing − original draft, Conceptualization. **Elizabeth R. Blackwood:** Writing − review & editing, Writing − original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Christopher Holland:** Writing − review & editing, Writing − original draft, Conceptualization. **Oren Gottfried:** Writing − review & editing, Conceptualization. **Christopher I. Shaffrey:** Writing − review & editing. **Muhammad M. Abd-El-Barr:** Writing − review & editing, Writing − original draft, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Supplementary materials

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.spinee.2024.09.010.

References

[1] Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990−2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016;388:1545–602. https://doi.org/10.1016/S0140-6736(16)31678-6.

[2] GBD 2021 Low Back Pain Collaborators. Global, regional, and national burden of low back pain, 1990-2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023;5:e316–29. https://doi.org/10.1016/S2665-9913(23)00098-X.

[3] Dieleman JL, Cao J, Chapin A, Chen C, Li Z, Liu A, et al. US Health Care spending by payer and health condition, 1996-2016. JAMA 2020;323:863–84. https://doi.org/10.1001/jama.2020.0734.

[4] Hoy D, March L, Brooks P, Blyth F, Woolf A, Bain C, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. Ann Rheumat Dis 2014;73:968–74. https://doi.org/10.1136/annrheumdis-2013-204428.

[5] Fatoye F, Gebrye T, Mbada CE, Useh U. Clinical and economic burden of low back pain in low- and middle-income countries: a systematic review. BMJ Open 2023;13:e064119. https://doi.org/10.1136/bmjopen-2022-064119.

[6] Han CS, Hancock MJ, Downie A, Jarvik JG, Koes BW, Machado GC, et al. Red flags to screen for vertebral fracture in people presenting with low back pain. Cochrane Database Syst Rev 2023;8: CD014461. https://doi.org/10.1002/14651858.CD014461.pub2.

[7] Mauck MC, Aylward AF, Barton CE, Birckhead B, Carey T, Dalton DM, et al. Evidence-based interventions to treat chronic low back pain: treatment selection for a personalized medicine approach. Pain Rep 2022;7:e1019. https://doi.org/10.1097/PR9.0000000000001019.

[8] Shipton EA. Physical therapy approaches in the treatment of low back pain. Pain Ther 2018;7:127–37. https://doi.org/10.1007/s40122-018-0105-x.

[9] Ho EK-Y, Chen L, Simic M, Ashton-James CE, Comachio J, Wang DXM, et al. Psychological interventions for chronic, non-specific low back pain: systematic review with network meta-analysis. BMJ 2022;376:e067718. https://doi.org/10.1136/bmj-2021-067718.

[10] Jenkins HJ, Ferreira G, Downie A, Maher C, Buchbinder R, Hancock MJ. The available evidence on the effectiveness of 10 common approaches to the management of non-specific low back pain: an evidence map. Eur J Pain 2022;26:1399–411. https://doi.org/10.1002/ejp.1974.

[11] Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, et al. Prevention and treatment of low back pain: evidence, challenges, and promising directions. Lancet 2018;391:2368–83. https://doi.org/10.1016/S0140-6736(18)30489-6.

[12] Allen-Watts K, Sims AM, Buchanan TL, DeJesus DJB, Quinn TL, Buford TW, et al. Sociodemographic differences in pain medication usage and healthcare provider utilization among adults with chronic low back pain. Front Pain Res 2022;2:806310. https://doi.org/10.3389/fpain.2021.806310.

[13] Bejarano G, Vining R, Desai DP, Minchew J, Michael Guo H, Goertz C. Development of a low back pain care pathway in an academic hospital system: results of a consensus process. J Orthop Surg Res 2024;19:11. https://doi.org/10.1186/s13018-023-04492-z.

[14] Ivanova JI, Birnbaum HG, Schiller M, Kantor E, Johnstone BM, Swindle RW. Real-world practice patterns, health-care utilization, and costs in patients with low back pain: the long road to guideline-concordant care. Spine J 2011;11:622–32. https://doi.org/10.1016/j.spinee.2011.03.017.

[15] Qaseem A, Wilt TJ, McLean RM, Forciea MA, Clinical Guidelines Committee of the American College of Physicians, Denberg TD, et al. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American college of physicians. Ann Intern Med 2017;166:514–30. https://doi.org/10.7326/M16-2367.

[16] Bernstein DN, Merchan N, Fear K, Rubery PT, Mesfin A. Greater socioeconomic disadvantage is associated with worse symptom severity at initial presentation in patients seeking care for lumbar disc herniation. Spine 2021;46:464–71. https://doi.org/10.1097/BRS.0000000000003811.

[17] Koes BW, van Tulder M, Lin C-WC, Macedo LG, McAuley J, Maher C. An updated overview of clinical guidelines for the management of non-specific low back pain in primary care. Eur Spine J 2010;19:2075–94. https://doi.org/10.1007/s00586-010-1502-y.

[18] Linton SJ, Shaw WS. Impact of psychological factors in the experience of pain. Phys Ther 2011;91:700–11. https://doi.org/10.2522/ptj.20100330.

[19] WHO. WHO guideline for non-surgical management of chronic primary low back pain in adults in primary and community care settings. Geneva: World Health Organization; 2023.

[20] Mathieu J, Robert M-È, Châtillon C-É, Descarreaux M, Marchand A-A. Appropriateness of specialized care referrals for LBP: a cross-sectional analysis. Front Med (Lausanne) 2024;10:1292481. https://doi.org/10.3389/fmed.2023.1292481.

[21] Deis N, Findlay JM. Appropriateness of lumbar spine referrals to a neurosurgical service. Can J Neurol Sci 2010;37:843–8. https://doi.org/10.1017/s0317167100051544.

[22] Debono B, Sabatier P, Koudsie A, Buffenoir K, Hamel O. Managing spine surgery referrals: the consultation of neurosurgery and its nuances. Neurochirurgie 2017;63:267–72. https://doi.org/10.1016/j.neuchi.2017.05.003.

[23] Greenwood-Lee J, Jewett L, Woodhouse L, Marshall DA. A categorisation of problems and solutions to improve patient referrals from primary to specialty care. BMC Health Serv Res 2018;18:986. https://doi.org/10.1186/s12913-018-3745-y.

[24] Chou R, Deyo R, Friedly J, Skelly A, Weimer M, Fu R, et al. Systemic pharmacologic therapies for low back pain: a systematic review for an American college of physicians clinical practice guideline. Ann Intern Med 2017;166:480–92. https://doi.org/10.7326/M16-2458.

[25] Chou R, Deyo R, Friedly J, Skelly A, Hashimoto R, Weimer M, et al. Nonpharmacologic therapies for low back pain: a systematic review for an American college of physicians clinical practice guideline. Ann Intern Med 2017;166:493–505. https://doi.org/10.7326/M16-2459.

[26] Burton W, Salsbury SA, Goertz CM. Healthcare provider perspectives on integrating a comprehensive spine care model in an academic health system: a cross-sectional survey. BMC Health Serv Res 2024;24:125. https://doi.org/10.1186/s12913-024-10578-z.

[27] Staartjes VE, Stumpo V, Kernbach JM, Klukowska AM, Gadjradj PS, Schröder ML, et al. Machine learning in neurosurgery: a global survey. Acta Neurochir (Wien) 2020;162:3081–91. https://doi.org/10.1007/s00701-020-04532-1.

[28] Iqbal J, Jahangir K, Mashkoor Y, Sultana N, Mehmood D, Ashraf M, et al. The future of artificial intelligence in neurosurgery: a narrative review. Surg Neurol Int 2022;13:536. https://doi.org/10.25259/SNI_877_2022.

[29] Dundar TT, Yurtsever I, Pehlivanoglu MK, Yildiz U, Eker A, Demir MA, et al. Machine learning-based surgical planning for neurosurgery: artificial intelligent approaches to the cranium. Front Surg 2022;9:863633. PMID: 35574559; PMCID: PMC9099011. https://doi.org/10.3389/fsurg.2022.863633.

[30] Raju B, Jumah F, Ashraf O, Narayan V, Gupta G, Sun H, et al. Big data, machine learning, and artificial intelligence: a field guide for neurosurgeons. J Neurosurg 2020;1:1–11. https://doi.org/10.3171/2020.5.JNS201288.

[31] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169:467–73. https://doi.org/10.7326/M18-0850.

[32] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 2019;25:44–56. https://doi.org/10.1038/s41591-018-0300-7.

[33] Seal HL. Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. Biometrika 1967;54:1–24. https://doi.org/10.1093/biomet/54.1-2.1.

[34] Tolles J, Meurer WJ. Logistic Regression: Relating Patient Characteristics to Outcomes. JAMA 2016;316:533. https://doi.org/10.1001/jama.2016.7653.

[35] Von Winterfeldt D, Edwards W. Decision analysis and behavioral research. Cambridge [Cambridgeshire]; New York: Cambridge University Press; 1986.

[36] Ho Tin Kam. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Machine Intell 1998;20:832–44. https://doi.org/10.1109/34.709601.

[37] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[38] Vapnik VN. The Support Vector method. In: Gerstner W, Germond A, Hasler M, Nicoud J-D, editors. Artificial neural networks — ICANN'97. Berlin, Heidelberg: Springer Berlin Heidelberg; 1997. p. 261–71 vol. 1327. https://doi.org/10.1007/BFb0020166.

[39] Lowd D, Domingos P. Naive Bayes models for probability estimation. In: Proceedings of the 22nd international conference on Machine learning - ICML '05. Bonn, Germany. ACM Press; 2005. p. 529–36. https://doi.org/10.1145/1102351.1102418.

[40] Gao CX, Dwyer D, Zhu Y, Smith CL, Du L, Filia KM, et al. An overview of clustering methods with guidelines for application in mental health research. Psychiat Res 2023;327:115265. https://doi.org/10.1016/j.psychres.2023.115265.

[41] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theo 1967;13:21–7. https://doi.org/10.1109/TIT.1967.1053964.

[42] Guida G, Mauri G. Evaluation of natural language processing systems: issues and approaches. Proc IEEE 1986;74:1026–35. https://doi.org/10.1109/PROC.1986.13580.

[43] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[44] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. Cambridge, MA: MIT Press; 2012.

[45] Stone M. Cross-validatory choice and assessment of statistical predictions. J Royal Stat Soc Series B: Stat Methodol 1974;36:111–33. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

[46] Larsen J, Hansen LK, Svarer C, Ohlsson M. Design and regularization of neural networks: the optimal use of a validation set. In: Neural Networks for Signal Processing VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop. Kyoto, Japan. IEEE; 1996. p. 62–71. https://doi.org/10.1109/NNSP.1996.548336.

[47] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.

[48] Reback J, jbrockmendel, McKinney W, Bossche JV den, Augspurger T, Roeschke M, et al. pandas-dev/pandas: Pandas 1.4.2 2022. https://doi.org/10.5281/zenodo.6408044.

[49] Caswell TA, Droettboom M, Lee A, Andrade ES de, Hoffmann T, Hunter J, et al. matplotlib/matplotlib: REL: v3.5.1 2021. https://doi.org/10.5281/zenodo.5773480.

[50] Waskom M, Gelbart M, Botvinnik O, Ostblom J, Hobson P, Lukauskas S, et al. mwaskom/seaborn: v0.12.2 2022. https://doi.org/10.5281/zenodo.7495530.

[51] Bishop JB, Szpalski M, Ananthraman SK, McIntyre DR, Pope MH. Classification of low back pain from dynamic motion characteristics using an artificial neural network. Spine (Phila Pa 1976) 1997;22:2991–8. https://doi.org/10.1097/00007632-199712150-00024.

[52] Bounds Lloyd, Mathew Waddell. A multilayer perceptron network for the diagnosis of low back pain. IEEE 1988 International Conference on Neural Networks; 1988. p. 481–9 vol. 2. https://doi.org/10.1109/ICNN.1988.23963.

[53] Gioftsos G, Grieve D. The use of artificial neural networks to identify patients with chronic low-back pain conditions from patterns of sit-to-stand manoeuvres. Clin Biomech 1996;11:275–80. https://doi.org/10.1016/0268-0033(96)00013-7.

[54] Mathew B, Norris D, Mackintosh I, Waddell G. Artificial intelligence in the prediction of operative findings in low back surgery. Br J Neurosurg 1989;3:161–70. https://doi.org/10.3109/02688698909002791.

[55] Oliver C, Atsma W. Artificial intelligence analysis of paraspinal power spectra. Clin Biomech 1996;11:422–4. https://doi.org/10.1016/0268-0033(96)00016-2.

[56] Sanders NW, Mann H. A comparative study of methods of processing patient pain drawings for analysis by an artificial neural network. In: Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society, vol. 1; 1995. p. 845–6 vol. 1. https://doi.org/10.1109/IEMBS.1995.575392.

[57] Vaughn ML, Cavill SJ, Taylor SJ, Foy MA, Fogg AJB. Using direct explanations to validate a multi-layer perceptron network that classifies low back pain patients. ICONIP'99. ANZIIS'99 & ANNES'99 & ACNN'99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378), vol. 2; 1999. p. 692–9 vol. 2. https://doi.org/10.1109/ICONIP.1999.845679.

[58] Zurada J, Karwowski W, Marras WS. A neural network-based system for classification of industrial jobs with respect to risk of low back disorders due to workplace design. Appl Ergonom 1997;28:49–58. https://doi.org/10.1016/S0003-6870(96)00034-8.

[59] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Machine Learn Res 2011;12:2825–30.

[60] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems 2016. https://doi.org/10.48550/arXiv.1603.04467.

[61] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library 2019. https://doi.org/10.48550/arXiv.1912.01703.

[62] SenthilKumar G, Madhusudhana S, Flitcroft M, Sheriff S, Thalji S, Merrill J, et al. Automated machine learning (AutoML) can predict 90-day mortality after gastrectomy for cancer. Sci Rep 2023;13:11051. https://doi.org/10.1038/s41598-023-37396-3.

[63] Musigmann M, Akkurt BH, Krähling H, Nacul NG, Remonda L, Sartoretti T, et al. Testing the applicability and performance of Auto ML for potential applications in diagnostic neuroradiology. Sci Rep 2022;12:13648. https://doi.org/10.1038/s41598-022-18028-8.

[64] Thirunavukarasu AJ, Elangovan K, Gutierrez L, Li Y, Tan I, Keane PA, et al. Democratizing artificial intelligence imaging analysis with automated machine learning: tutorial. J Med Intern Res 2023;25:e49949. https://doi.org/10.2196/49949.

[65] Sundberg L, Holmström J. Democratizing artificial intelligence: how no-code AI can leverage machine learning operations. Business Horizons 2023;66:777–88. https://doi.org/10.1016/j.bushor.2023.04.003.

[66] Gao S, Han L, Luo D, Liu G, Xiao Z, Shan G, et al. Modeling drug mechanism of action with large scale gene-expression profiles using GPAR, an artificial intelligence platform. BMC Bioinformat 2021;22:17. https://doi.org/10.1186/s12859-020-03915-6.

[67] Mahmoud SS, Pallaud RF, Kumar A, Faisal S, Wang Y, Fang Q. A comparative investigation of automatic speech recognition platforms for aphasia assessment batteries. Sensors (Basel) 2023;23:857. https://doi.org/10.3390/s23020857.

[68] United States: National Archives and Records Administration: Office of the Federal Register, United States: Department of Health and Human Services, United States: Centers for Medicare & Medicaid Services. Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Policy Changes and Fiscal Year 2019 Rates; Quality Reporting Requirements for Specific Providers; Medicare and Medicaid Electronic Health Record (EHR) Incentive Programs (Promoting Interoperability Programs) Requirements for

Eligible Hospitals, Critical Access Hospitals, and Eligible Professionals; Medicare Cost Reporting Requirements; and Physician Certification and Recertification of Claims. Part II: Rules and Regulations. Office of the Federal Register, National Archives and Records Administration; 2018.

[69] Anumula N, Sanelli PC. Meaningful use. AJNR Am J Neuroradiol 2012;33:1455–7. https://doi.org/10.3174/ajnr.A3247.

[70] National trends in hospital and physician adoption of electronic health records | HealthIT.gov n.d. https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records. (Accessed March 26, 2024).

[71] Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023;619:357–62. https://doi.org/10.1038/s41586-023-06160-y.

[72] Pugliese R, Regondi S, Marini R. Machine learning-based approach: global trends, research directions, and regulatory standpoints. Data Sci Manage 2021;4:19–29. https://doi.org/10.1016/j.dsm.2021.12.002.

[73] Kohane IS. Injecting artificial intelligence into medicine. NEJM AI 2023;1:AIe2300197. https://doi.org/10.1056/AIe2300197.

[74] Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artific Intell Healthcare 2020:25–60. https://doi.org/10.1016/B978-0-12-818438-7.00002-2. Epub 2020 Jun 26. PMCID: PMC7325854.

[75] Ozaydin B, Berner ES, Cimino JJ. Appropriate use of machine learning in healthcare. Intell Based Med 2021;5:100041. https://doi.org/10.1016/j.ibmed.2021.100041.

[76] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26:1364–74. https://doi.org/10.1038/s41591-020-1034-x.

[77] Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, et al. SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1038/s41591-020-1037-7.

[78] Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, Young A, Jelovsek JE, O'Brien C, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc 2022;29:1631–6. https://doi.org/10.1093/jamia/ocac078.

[79] Abdollah V, Parent EC, Dolatabadi S, Marr E, Croutze R, Wachowicz K, et al. Texture analysis in the classification of T2-weighted magnetic resonance images in persons with and without low back pain. J Orthopaed Res 2021;39:2187–96. https://doi.org/10.1002/jor.24930.

[80] Agarwal N, Aabedi AA, Chan AK, Letchuman V, Shabani S, Bisson EF, et al. Leveraging machine learning to ascertain the implications of preoperative body mass index on surgical outcomes for 282 patients with preoperative obesity and lumbar spondylolisthesis in the Quality Outcomes Database. J Neurosurg Spine 2023;38:182–91. https://doi.org/10.3171/2022.8.SPINE22365.

[81] Gross DP, Zhang J, Steenstra I, Barnsley S, Haws C, Amell T, et al. Development of a computer-based clinical decision support tool for selecting appropriate rehabilitation interventions for injured workers. J Occup Rehabil 2013;23:597–609. https://doi.org/10.1007/s10926-013-9430-4.

[82] Krebs B, Nataraj A, McCabe E, Clark S, Sufiyan Z, Yamamoto SS, et al. Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports. Eur Spine J 2023;161(3):345–56. https://doi.org/10.1007/s00586-023-07552-4.

[83] Naumzik C, Feuerriegel S, Nielsen AM. Data-driven dynamic treatment planning for chronic diseases. Eur J Operat Res 2023;305:853–67. https://doi.org/10.1016/j.ejor.2022.06.016.

[84] Oude Nijeweme-d'Hollosy W, van Velsen L, Poel M, Groothuis-Oudshoorn CGM, Soer R, Hermens H. Evaluation of three machine learning models for self-referral decision support on low back pain in primary care. Int J Med Inform 2018;110:31–41. https://doi.org/10.1016/j.ijmedinf.2017.11.010.

[85] d'Hollosy WON-, van Velsen L, Poel M, Groothuis-Oudshoorn C, Soer R, Stegeman P, et al. Applying machine learning on patient-reported data to model the selection of appropriate treatments for low back pain: a pilot study. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020). SCITEPRESS; 2020. p. 117–24. https://doi.org/10.5220/0008962101170124.

[86] Purohit G, Choudhary M, Sinha VD. Use of Artificial Intelligence for the development of predictive model to help in decision-making for patients with degenerative lumbar spine disease. Asian J Neurosurg 2022;17:274–9. https://doi.org/10.1055/s-0042-1750785.

[87] Knoop J, van Lankveld W, Beijer L, Geerdink FJB, Heymans MW, Hoogeboom TJ, et al. Development and internal validation of a machine learning prediction model for low back pain non-recovery in patients with an acute episode consulting a physiotherapist in primary care. BMC Musculoskelet Disord 2022;23:834. https://doi.org/10.1186/s12891-022-05718-7.

[88] Liu P, Wu Y, Xiao Z, Gold LS, Heagerty PJ, Annaswamy T, et al. Estimating individualized treatment effects using a risk-modeling approach: an application to epidural steroid injections for lumbar spinal stenosis. Pain 2023;164:811–9. https://doi.org/10.1097/j.pain.0000000000002768.

[89] Liew BXW, Ford JJ, Scutari M, Hahne AJ. How does individualised physiotherapy work for people with low back pain? A Bayesian Network analysis using randomised controlled trial data. PLoS One 2021;16:e0258515. https://doi.org/10.1371/journal.pone.0258515.

[90] Eisenberg DM, Buring JE, Hrbek AL, Davis RB, Connelly MT, Cherkin DC, et al. A model of integrative care for low-back pain. J Altern Complement Med 2012;18:354–62. https://doi.org/10.1089/acm.2011.0408.

[91] Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. In: 2009 Ninth IEEE International Conference on Data Mining. Miami Beach, FL, USA. IEEE; 2009. p. 908–13. https://doi.org/10.1109/ICDM.2009.108.

[92] Hamilton DG, Hong K, Fraser H, Rowhani-Farid A, Fidler F, Page MJ. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. BMJ 2023;382:e075767. https://doi.org/10.1136/bmj-2023-075767.

[93] Venkatesh K, Santomartino SM, Sulam J, Yi PH. Code and data sharing practices in the radiology artificial intelligence literature: a meta-research study. Radiolo Artific Intell 2022;4:e220081. https://doi.org/10.1148/ryai.220081.

[94] Kazerooni AF, Khalili N, Liu X, Haldar D, Jiang Z, Anwar SM, et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs) 2024. https://doi.org/10.48550/arXiv.2305.17033.

[95] RSNA Abdominal Trauma Detection AI Challenge (2023) n.d. https://www.rsna.org/rsnai/ai-image-challenge/abdominal-trauma-detection-ai-challenge. Accessed March 26, 2024.

[96] Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. J Med Internet Res 2023;25:e48763. https://doi.org/10.2196/48763.