Review

Clinical decision support systems for the management of low back pain: A systematic review

Elisa Caria[1], Giulia Delrio[1], Ilaria Pinna[1], Salvatore Sardu[1], Corrado Ciatti[2,3], Francesco Muresu[1], Fabio Milia[1], Gianfilippo Caggiari[1], Carlo Doria[1]

[1]Orthopaedic Department, Sassari University Hospital, Sassari, Italy; [2]Department of Orthopaedics and Traumatology, Guglielmo da Saliceto Hospital, Piacenza, Italy; [3]University of Parma, Parma, Italy

Abstract. *Background and aim:* Low back pain (LBP) is one of the most prevalent musculoskeletal disorders and a major contributor to years lived with disability worldwide. Despite international guidelines promoting early conservative management and rational imaging use, clinical practice remains highly variable. Clinical Decision Support Systems (CDSS), particularly those integrating artificial intelligence (AI), have been developed to support diagnostic accuracy and standardize care. This review aims to synthesize current evidence on the performance, applications, and clinical integration of CDSS in the management of LBP. *Methods:* A systematic review was conducted according to PRISMA 2020 guidelines. Searches were performed in PubMed, Scopus, ProQuest, and PsycINFO up to July 2025, including original studies evaluating CDSS for diagnosis or treatment planning in adults with LBP. Data extraction covered study design, CDSS type, data sources, performance metrics, and clinical outcomes. Risk of bias was assessed using QUADAS-2/QUADAS-AI, RoB 2, and ROBINS-I tools as appropriate. Descriptive statistics were computed for accuracy, sensitivity, specificity, and area under the curve (AUC). *Results:* Nineteen studies met the inclusion criteria. Mean diagnostic accuracy was 0.911 (median 0.916), with corresponding mean sensitivity, specificity, and AUC of 0.865, 0.896, and 0.830, respectively. AI-based and hybrid systems performed comparably to rule-based models, while imaging optimization studies showed reductions of approximately 10% in unnecessary imaging and 15% in MRI utilization. *Conclusions:* CDSS demonstrate high diagnostic performance and potential to improve guideline adherence and resource efficiency in LBP care. Broader implementation requires evaluation of long-term patient outcomes, cost-effectiveness, and real-world integration within electronic health records. (www.actabiomedica.it)

Key words: low back pain, low back pain management, clinical decisional support system, systematic review

Introduction

Low back pain (LBP) is one of the most widespread musculoskeletal disorders worldwide, with lifetime prevalence estimates indicating that up to 80% of individuals will experience at least one episode during their lives (1-9). The burden is particularly high among working-age adults, but older populations also contribute significantly to healthcare utilization (8). LBP is a leading cause of years lived with disability (YLDs), as reported by the Global Burden of Disease Study, and represents a substantial socioeconomic challenge due to direct healthcare costs and indirect costs related to work absenteeism and reduced productivity (1-9). The etiology of LBP is heterogeneous, encompassing mechanical causes, degenerative changes, and, in a minority of cases, serious underlying pathologies such as fractures, infections, or malignancies (10). Clinical presentations vary from acute self-limiting episodes to chronic, disabling pain, often influenced by psychosocial and occupational factors (11,12). Given this complexity, optimal management frequently requires a multidisciplinary approach involving primary care physicians, physiotherapists, orthopaedic specialists, radiologists, and mental health professionals (13). Despite the existence of international guidelines promoting conservative management, timely identification of red flags, and avoidance of unnecessary imaging (14-16), adherence in clinical practice remains inconsistent (17,18). Variations in clinician expertise, diagnostic uncertainty, patient expectations, and differences in resource availability contribute to heterogeneous care pathways. Such variability can lead to overuse of imaging, delayed initiation of evidence-based therapies, and, in some cases, unnecessary invasive procedures. Clinical Decision Support Systems (CDSS) have emerged as a promising technological solution to these challenges. CDSS are designed to integrate patient-specific data with structured clinical knowledge, providing clinicians with evidence-based recommendations at the point of care (19). AI-based CDSS harness machine learning and deep learning techniques to detect patterns in complex datasets, including medical imaging and electronic health records, while rule-based systems translate guideline recommendations into algorithmic pathways (20). Hybrid models combine AI's adaptability with the interpretability of rule-based logic, offering a balance between innovation and clinical transparency (21). In LBP management, CDSS can assist in early detection of red flags, guide risk stratification for chronicity, support diagnostic decision-making, and recommend appropriate treatment strategies aligned with guidelines (22,23). These systems have the potential to standardize care, reduce unwarranted variation, and improve resource allocation. This review synthesizes evidence from 19 original studies assessing the performance, application, and clinical integration of CDSS for LBP, with a statistical overview of diagnostic accuracy and decision-support metrics.

Material and Methods

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines, ensuring transparency and reproducibility in every phase of the process. The protocol was designed a priori and included a structured search strategy, predefined eligibility criteria, independent screening by multiple reviewers, and a rigorous approach to data extraction and synthesis. A comprehensive literature search was carried out in PubMed, Scopus, ProQuest, and PsycINFO, covering all publications up to July 14, 2025. The search strategy combined both controlled vocabulary terms (e.g., MeSH and Emtree headings) and free-text keywords related to Clinical Decision Support Systems and Low Back Pain. The exact search strings were adapted to the syntax of each database, and the complete PubMed search query is reported in Supplementary Material A. No language or date restrictions were applied to ensure a comprehensive evidence base. All retrieved citations were imported into Zotero reference management software for deduplication. Two reviewers independently screened the titles and abstracts of all records to assess their relevance. Articles that appeared to meet the inclusion criteria, or where eligibility was uncertain, underwent full-text review. Any discrepancies between reviewers were resolved through discussion, and, if necessary, adjudicated by a third reviewer. Studies were included if they met the following criteria:

1. Design – Original research articles, either prospective or retrospective.
2. Population – Adult patients (≥18 years) with low back pain.
3. Intervention – Use of a CDSS in diagnosis, clinical decision-making, or treatment planning for LBP.
4. Outcomes – Reporting of at least one quantitative metric (accuracy, sensitivity, specificity, or area under the ROC curve) or qualitative outcome related to clinical impact or implementation feasibility.

Exclusion criteria were: narrative reviews, systematic reviews, conference abstracts without full text, study protocols, purely algorithmic studies without clinical validation, and retracted papers.

Data extraction

A standardized data extraction form was used to collect information from each study, including:

- Authors and year of publication
- Country and clinical setting
- Study design and sample size
- CDSS type (AI-based, rule-based, hybrid)
- Input data sources (e.g., imaging, electronic health records, patient-reported outcomes)
- Reported performance metrics (accuracy, sensitivity, specificity, AUC)
- Key clinical and implementation findings

Supplementary Material B provides a detailed summary of the characteristics of the included studies.

Risk of bias assessment

Two reviewers independently evaluated the risk of bias using design-appropriate tools: QUADAS-2/QUADAS-AI for diagnostic accuracy and model development/validation studies, RoB 2 for randomized trials, and ROBINS-I for non-randomized studies. Disagreements were resolved by consensus or a third reviewer. Domain-level and overall judgments are provided in Supplementary Material B (Table S2–S3; Figure S1). These judgments informed the qualitative synthesis and interpretation of results.

Statistical analysis

Descriptive statistics (mean, standard deviation, range) were computed for the main performance metrics: accuracy, sensitivity, specificity, and AUC. Visualizations included boxplots with swarm overlays to display both the distribution and individual study values, and bar charts with 95% confidence intervals to compare mean performance across metrics. Statistical analyses were performed using IBM SPSS Statistics for Windows, Version 29.0 (IBM Corp., Armonk, NY, USA). Missing outcome data were addressed using case-wise deletion, ensuring that each metric's analysis was based only on studies that reported that metric. We summarised performance per metric using the mean, median and inter-quartile range (IQR), and also computed a sample-size–weighted mean (weights = study N when reported); no meta-analysis was performed due to outcome heterogeneity.

Results

The database search retrieved 2,300 records in total (PubMed 287, Scopus 1,715, ProQuest 250, PsycINFO 48) (Figure 1). After removing 374 duplicates, 1,926 titles and abstracts were screened. Fifty-five full texts were assessed for eligibility; two could not be retrieved. Thirty-four articles were excluded for pre-specified reasons, leaving 19 studies for qualitative synthesis.

The 19 studies encompass a broad range of designs and contexts.

Risk of bias across studies.

The overall risk of bias varied by design. In diagnostic accuracy and development/validation studies, the most frequent concerns related to patient selection, blinding of index tests/reference standards, and lack of external validation. In randomized and non-randomized evaluations, common issues involved allocation concealment, deviations from intended interventions, and selective reporting. Full domain-level judgments and justifications are presented in Table S3 and Figure S1 (Supplementary Material B). When grouped by design, seven were prospective or observational evaluations (including pilot or multicenter cross-sectional work), seven were development/validation studies focused on algorithmic performance, and four were randomized or cluster-randomized trials; one article was classified as other. The geographical distribution was heterogeneous; the most represented settings were the USA (n=2) (24,25), Germany (n=2) (26,27), and Iran (n=2) (28,29), with single studies from Japan (30), and a range of other single-country studies across Europe, Asia, and the Middle East, as well as multicountry collaborations (e.g., USA–Europe, Nordic countries, and global consortia). In terms of technology, the studies split into rule-based CDSS (n=6) (24,26,27,31-33), AI-based systems (n=7) (28), (34-39),

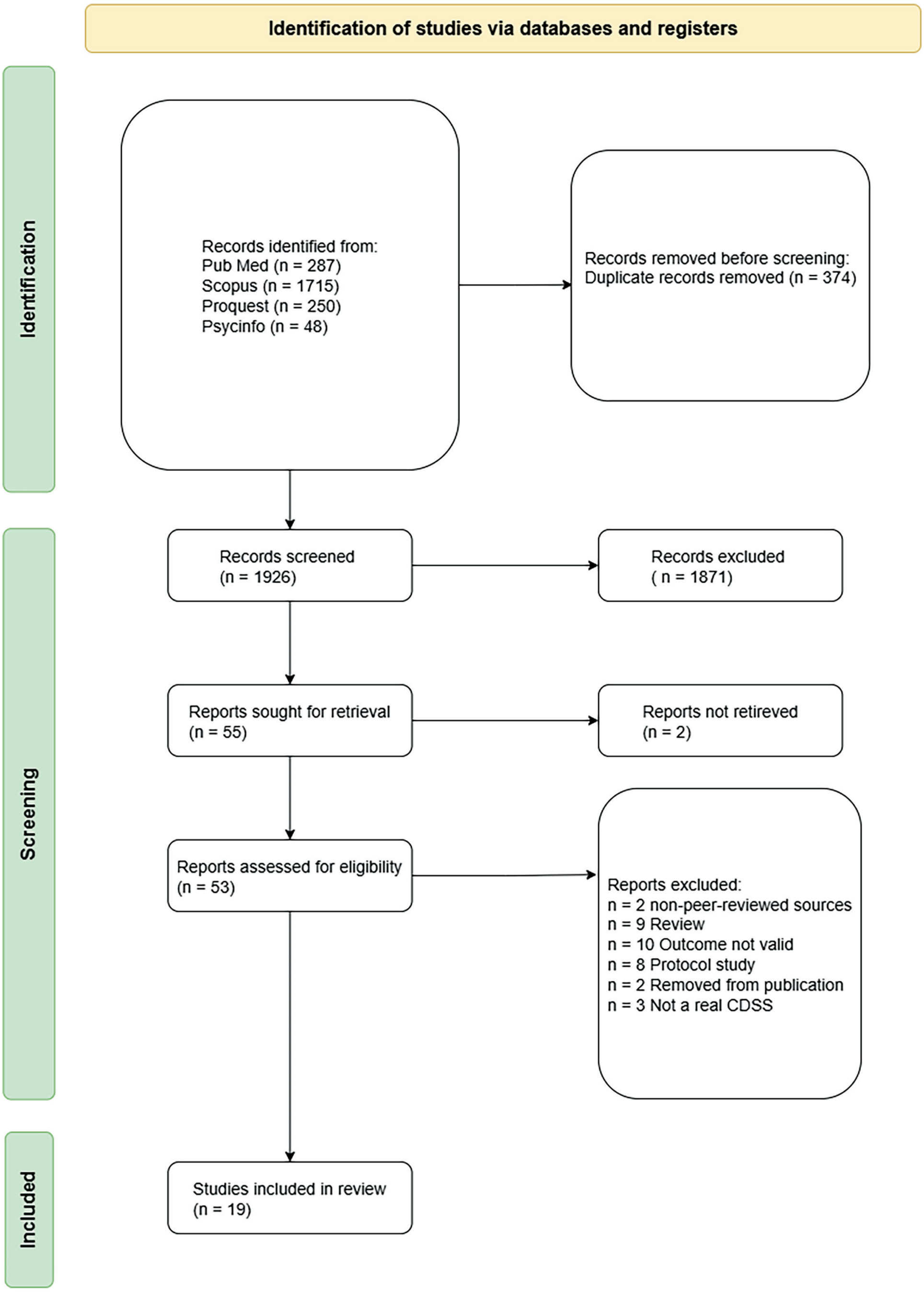

Figure 1. PRISMA 2020 flow diagram illustrating the identification, screening, and inclusion of studies assessing Clinical Decision Support Systems (CDSS) for low back pain.

and hybrid approaches (n=6) (25,29,30), (40-42) (Figure 2).

With respect to input data, clinical/EHR-driven systems were predominant (n=15) (24-28), (30), (32–34), (36), (38-42), followed by imaging-based systems (n=2) (35,37) and other/biotech or mixed modalities (n=2) (29,31) (Figure 3). Mapped to functional intent, imaging optimization (appropriateness and reporting automation) was the commonest target (9 studies (24-27,31,33,35,36,42)), followed

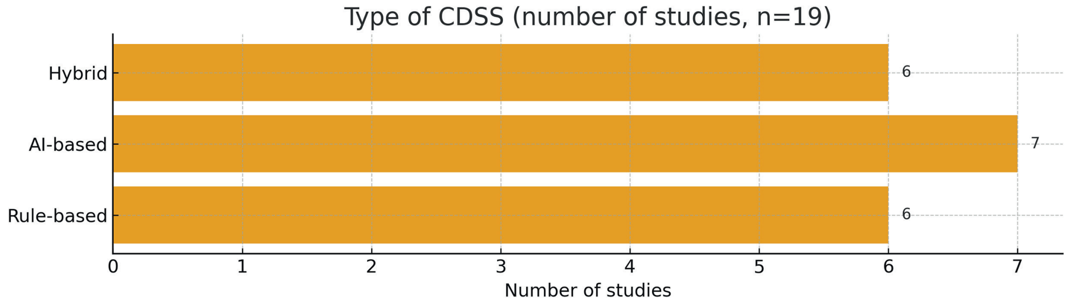

Figure 2. Distribution of included studies by CDSS type. Among the 19 studies, 7 used AI-based systems, 6 rule-based systems, and 6 hybrid models integrating artificial intelligence with rule-based logic.

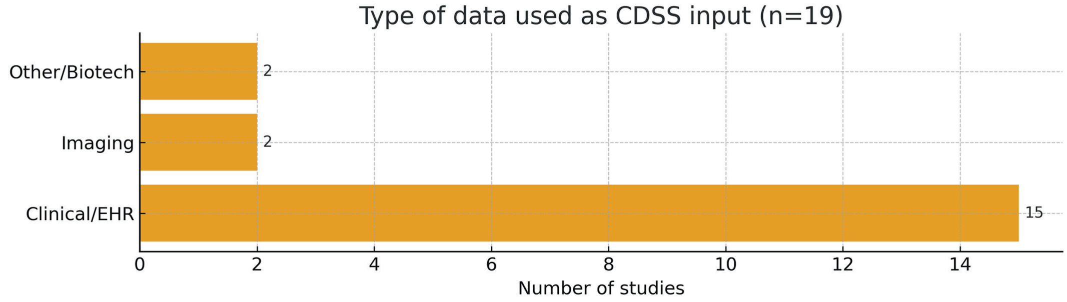

Figure 3. Type of data used as input for CDSS. Most systems (n=15) relied on clinical or electronic health record (EHR) data, whereas imaging-based and other/biotechnological inputs were less common (n=2 each).

by diagnosis/severity support (6 studies (28,29,37-39,41)) and triage/screening (2 studies (30, 32)); two studies addressed other decision needs (34,40). Performance across studies. For each metric we report the mean, median and inter-quartile range (IQR), and a sample-size–weighted mean based on the subset of studies that provided the metric. Accuracy: mean 0.911, median 0.916, IQR 0.890–0.937, weighted mean 0.915. Sensitivity: mean 0.865, median 0.870, IQR 0.835–0.893, weighted mean 0.867. Specificity: mean 0.896, median 0.900, IQR 0.880–0.930, weighted mean 0.902. AUC: mean 0.830, median 0.830, IQR 0.820–0.845, weighted mean 0.844. These summary statistics correspond to Figure 4 (IQR = line; median = dot; mean = X; weighted mean = square).

Only a minority of studies reported patient-centred outcomes. Among those, one study quantified pain reduction with an average effect size of 0.32 (standardized), while disability change was rarely reported in a way that allowed synthesis. Implementation-level outcomes were more common within imaging-optimization trials: in the before/after and cluster-randomized comparisons that reported explicit figures, the overall imaging rate decreased by ~9.6%, and MRI utilization fell by ~14.9%, with one health-system analysis estimating $1.87 million in cost savings over the evaluation horizon. Although these results derive from a small subset, they suggest that CDSS can materially influence ordering behavior and downstream costs when embedded into workflows. Qualitative findings were consistent with the quantitative signal. Imaging-based AI systems, particularly those using convolutional architectures, performed well for automated reading tasks and structured reporting, achieving accuracies in the upper-80s to mid-90s while standardizing outputs across observers. Rule-based tools, typically encoding guideline logic for low back pain, improved appropriateness of imaging and referrals and provided

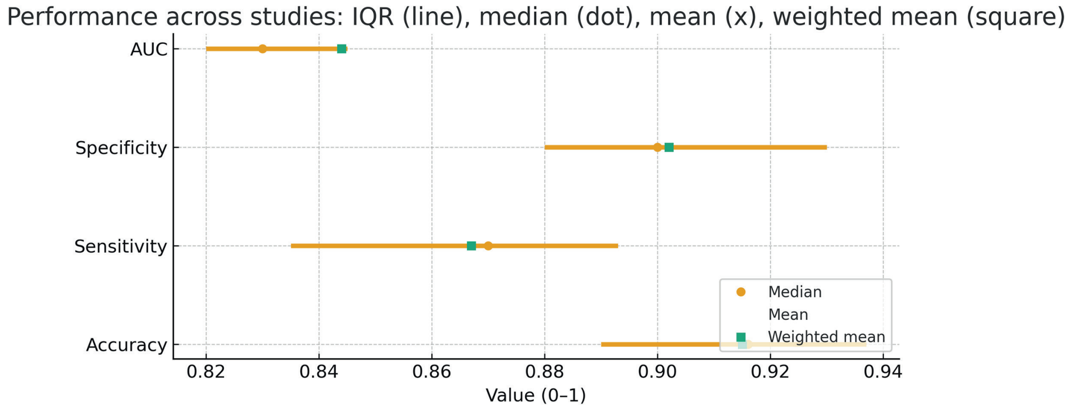

Figure 4. Summary of performance across studies. Median, mean, and weighted mean values (accuracy, sensitivity, specificity, and AUC) are presented with interquartile ranges, showing overall high diagnostic performance of CDSS in low back pain management.

transparent, auditable rationales that clinicians valued. Hybrid systems balanced these strengths, using AI for feature discovery while constraining recommendations within explicit rules, which appeared to increase clinical acceptance. In sensitivity analyses limited to studies using clinical/EHR inputs rather than imaging, performance remained robust, with median accuracies in the low-to-mid 0.90s and no systematic decrement in sensitivity relative to imaging-first systems. However, heterogeneity in case-mix, endpoints, and validation design (retrospective vs. prospective) precluded meta-analysis beyond descriptive pooling. Overall, the pattern across metrics, designs, and targets supports the conclusion that CDSS for low back pain deliver high accuracy (≈0.91), balanced sensitivity/specificity (≈0.86/0.90), and solid AUC (≈0.83) in real datasets, while early implementation studies signal reductions in unnecessary imaging and potential cost savings, albeit with limited evidence to date on long-term patient outcomes.

Discussion

The findings of this systematic review highlight the increasing role of Clinical Decision Support Systems (CDSS) in the management of low back pain (LBP), demonstrating consistent diagnostic accuracy and early signals of clinical utility. Across diverse methodologies, CDSS achieved a pooled mean accuracy of 0.91 and balanced sensitivity/specificity around 0.86/0.90, confirming their ability to provide reliable assistance in clinical decision-making. These results align with prior reviews on decision support tools in musculoskeletal care, which have emphasized their potential to reduce diagnostic variability and improve adherence to evidence-based guidelines (19,22).

Interpretation of findings

Rule-based systems showed particular strength in standardizing guideline adherence and reducing unnecessary imaging, especially when embedded into electronic health record (EHR) workflows. Their interpretability and transparent logic appear to foster clinician trust—an essential factor for adoption. Conversely, AI-based systems, particularly those using deep learning on imaging data, demonstrated superior pattern recognition and automation capabilities but often lacked external validation and interpretability. Hybrid models combining both paradigms offered the most promising balance between performance and explainability, suggesting that future systems should continue integrating explicit rule constraints with data-driven intelligence (21). Although diagnostic accuracy was consistently high, the translation into improved patient outcomes remains limited. Only a few studies reported patient-centred metrics such as pain reduction or disability improvement, and even those were short-term or exploratory. This evidentiary gap mirrors broader trends in digital health research, where technical validation precedes large-scale pragmatic evaluation. The moderate-to-high risk of bias observed in many development studies—especially regarding patient selection, absence of blinding, and internal-only validation—further underscores the need for robust, externally validated prospective trials (17).

Implementation and clinical impact

Implementation-focused studies, primarily in radiology and primary care, indicate that CDSS can meaningfully alter clinician behaviour. In the included before/after and cluster-randomized evaluations, imaging requests decreased by roughly 10–15%, accompanied by estimated system-level cost savings approaching two million USD. These data, although preliminary, support the view that CDSS can serve not only as diagnostic aids but also as instruments of resource stewardship. However, most evaluations occurred in high-income settings with mature digital infrastructures; generalizability to low- and middle-income contexts remains uncertain (9). Adoption barriers commonly reported across studies included workflow disruption, alert fatigue, lack of interoperability with legacy EHRs, and insufficient clinician training. Conversely, enablers included real-time feedback, transparent rationale for recommendations, and adaptability to local guidelines. These findings align with implementation science literature emphasizing usability, co-design, and organizational readiness as critical determinants of sustained CDSS use (6).

Comparison with prior evidence

Previous narrative and scoping reviews have documented variable results for decision support in musculoskeletal disorders, often limited by small samples or simulation-based designs (19,20). By including recent AI-driven systems and hybrid approaches, the present review provides an updated synthesis reflecting current methodological maturity. The performance metrics observed here are comparable to those in spine-specific AI reviews (22) and exceed those typically reported in general diagnostic support systems for primary care, where accuracies around 0.80 are common. This suggests that domain-specific knowledge integration and narrower clinical focus may yield superior performance.

Strengths and limitations

This review benefits from a comprehensive search across four databases without language restrictions, rigorous dual screening, and use of design-specific risk-of-bias tools (QUADAS-2/AI, RoB 2, ROBINS-I). Nonetheless, several limitations must be acknowledged. First, quantitative synthesis was descriptive due to heterogeneity in design, endpoints, and reported outcomes. Second, publication bias may exist, as negative or low-performance studies are less likely to be published. Third, performance reporting was inconsistent: key metrics such as AUC or calibration were frequently omitted, preventing full comparability. Finally, the majority of studies evaluated technical accuracy rather than real-world clinical effectiveness, limiting conclusions about long-term impact on patient care.

Implications for research and practice

The current evidence suggests that CDSS for LBP can reach a mature level of technical reliability and can positively influence clinician behaviour when integrated into workflow. However, to achieve broad clinical impact, future research should move beyond algorithmic validation toward pragmatic trials evaluating long-term patient outcomes, cost-effectiveness, and equity of access. Integration with interoperable EHRs, explainable AI techniques, and continuous feedback loops will be essential to ensure clinician confidence and patient safety. Finally, multidisciplinary collaboration between clinicians, data scientists, and human-factors experts will be critical to bridge the gap between algorithmic performance and meaningful clinical benefit.

Conclusion

Clinical Decision Support Systems (CDSS) for the management of low back pain (LBP) have emerged as reliable and adaptable tools capable of enhancing clinical decision-making in a variety of settings. Their consistent high performance across diverse methodologies and data sources suggests that they can play a significant role in improving diagnostic precision, supporting adherence to guidelines, and fostering a more standardized approach to patient care. By integrating evidence-based knowledge with patient-specific information, CDSS can assist clinicians in identifying critical diagnostic cues, optimizing the use of imaging, and guiding appropriate referrals. These systems also offer potential benefits in terms of workflow efficiency and resource allocation, reducing unnecessary investigations and facilitating timely interventions. Nevertheless, their successful adoption in routine practice depends on more than technical accuracy. Long-term patient outcomes, cost-effectiveness, and seamless integration into existing electronic health record systems remain areas requiring further exploration. Equally important are the human factors: clinician training, user acceptance, and the transparency of decision-making processes all influence the real-world utility of these technologies. Future research should focus on evaluating CDSS in real-world clinical environments through longitudinal and pragmatic studies, assessing their impact not only on diagnostic outcomes but also on patient-reported measures and health system performance. Furthermore, considerations of equity, accessibility, and ethical AI deployment will be essential to ensure that these innovations contribute to high-quality, patient-centred care on a broad scale. With sustained research, thoughtful implementation, and attention to the needs of both clinicians and patients, CDSS have the potential to become integral tools in the comprehensive management of LBP.

Ethic Approval: Ethical approval was not required for this study, as it is a systematic review of previously published data and did not involve human participants or identifiable personal information.

Conflict of Interest: Each author declares that he or she has no commercial associations (e.g. consultancies, stock ownership, equity interest, patent/licensing arrangement etc.) that might pose a conflict of interest in connection with the submitted article.

Authors Contribution: EC, GD, IP, SS, GC: Conceptualization; EC, GC: Methodology; GD: Software; IP: Validation; SS: Formal analysis; FM: Investigation; GC: Resources; EC: Data curation; EC, GD, IP, SS, GC: Writing – original draft; CC, FM, GC, CD: Writing – review and editing; CC: Visualization; CD: Supervision. All authors have read and agreed to the published version of the manuscript.

Declaration on the Use of AI: The authors used artificial intelligence–assisted software (ChatGPT, OpenAI, San Francisco, CA, USA) exclusively for English language editing and grammar refinement. The final text was entirely reviewed and approved by the authors.

Acknowledgments: The authors gratefully acknowledge Prof. Gianluca Vadalà from Campus Bio-Medico University of Rome for facilitating access to the funding that made this research possible.

Funding: This research was funded by the European Union under the Next Generation EU iniziative – NRRP M6C2, Investment 2.1: Enhancement and Strengthening of Biomedical Research in the NHS (project number PNRR-MAD-2022-12376692, CUP H83C22000940001).

References

1. Ferreira ML, De Luca K, Haile LM, et al. 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(6):e316–e329. doi:10.1016/S2665-9913(23)00098-X
2. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71
3. Dijk SW, Wollny C, Barkhausen J, et al. Evaluation of a clinical decision support system for imaging requests: a cluster randomized clinical trial. JAMA. 2025;333:7439. doi:10.1001/jama.2024.27853
4. Tsega S, Krouss M, Alaiev D, et al. Imaging wisely campaign: initiative to reduce imaging for low back pain across a large safety net system. J Am Coll Radiol. 2024;21(1):165–174. doi:10.1016/j.jacr.2023.07.012
5. Chen R, Zhang P, Mao Z, et al. Artificial intelligence for lumbosacral degenerative spine disease: advancements, challenges, and future directions. World Neurosurg. 2020;134:e672–e682. doi:10.1016/j.wneu.2019.11.031
6. Preti LM, Ardito V, Compagni A, Petracca F, Cappellaro G. Implementation of machine learning applications in health care organizations: systematic review of empirical studies. J Med Internet Res. 2024;26:e55897. doi:10.2196/55897
7. Shokri P, Zahmatyar M, Falah Tafti M, et al. Non-spinal low back pain: global epidemiology, trends, and risk factors. Health Sci Rep. 2023;6(9):e1533. doi:10.1002/hsr2.1533

8. World Health Organization. Low back pain – fact sheet. Geneva: WHO; 2023. INCOMPLETE (website fields missing)
9. 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(4):e064119. doi:10.1136/bmjopen-2022-064119
10. Gushcha AO, Sharif S, Zileli M, Oertel J, Zygourakis CC, Yusupova AR. Acute back pain: clinical and radiologic diagnosis: WFNS spine committee recommendations. World Neurosurg X. 2024;22:100278. doi:10.1016/j.wnsx.2024.100278
11. Dunn M, Rushton AB, Mistry J, Soundy A, Heneghan NR. The biopsychosocial factors associated with development of chronic musculoskeletal pain: an umbrella review and meta-analysis of observational systematic reviews. PLoS One. 2024;19(4):e0294830. doi:10.1371/journal.pone.0294830
12. Jahn A, Andersen JH, Christiansen DH, Seidler A, Dalbøge A. Occupational mechanical exposures as risk factor for chronic low-back pain: a systematic review and meta-analysis. Scand J Work Environ Health. 2023;49(7):453–465. doi:10.5271/sjweh.4114
13. World Health Organization. (2023). WHO guideline for non-surgical management of chronic primary low back pain. Geneva: WHO. Organizzazione Mondiale della Sanità. Available from: https://www.who.int/publications/i/item/9789240081789
14. Low back pain and sciatica in over 16s: assessment and management. London: National Institute for Health and Care Excellence (NICE); 2020 Dec 11. (NICE Guideline, No. 59.) Available from: https://www.ncbi.nlm.nih.gov/books/NBK562933/
15. Nicol V, Verdaguer C, Daste C, et al. Chronic low back pain: a narrative review of recent international guidelines for diagnosis and conservative treatment. J Clin Med. 2023; 12(4):1685. doi:10.3390/jcm12041685
16. Hall AM, Kamper SJ, Maher CG. Do not routinely offer imaging for uncomplicated low back pain. BMJ. 2021;372:n291. doi:10.1136/bmj.n291
17. Samanna CL, Buntine P, Belavy DL, et al. Adherence to low back pain clinical guidelines in Australian hospital emergency departments: a public and private comparison. Australas Emerg Care. 2024;27(4):276–281. doi:10.1016/j.auec.2024.07.001
18. Heine J, Window P, Hacker S, et al. Adherence to recommended guidelines for low back pain presentations to an Australian emergency department: barriers and enablers. Australas Emerg Care. 2023;26(4):326–332. doi:10.1016/j.auec.2023.04.003
19. Toh ZA, Polat A, Liew D. Clinical decision support systems in spine care: a scoping review. JMIR Med Inform. 2024;12:e58122. doi:10.2196/58122
20. Tagliaferri SD, Angelova M, Zhao X, et al. Artificial intelligence to improve back pain outcomes: machine learning and beyond. NPJ Digit Med. 2020;3:93. doi:10.1038/s41746-020-0289-6
21. Kierner S, Kierner P, Kucharski J. Combining machine learning models and rule engines in clinical decision systems: exploring optimal aggregation methods for vaccine hesitancy prediction. Comput Biol Med. 2025;188:109749. doi:10.1016/j.compbiomed.2025.109749
22. Giaccone M, Ferraro S, Placci A, Giorgi Pierfranceschi M. AI-based decision support systems in lumbar degenerative spine disorders: a systematic review. BMC Musculoskelet Disord. 2025;26:85. doi:10.1186/s12891-025-08254-1
23. Rudin S, Sharma A, Agrawal A. A patient-centered clinical decision support tool for low back pain: usability study. JMIR Form Res. 2025;9:e58051. doi:10.2196/58051
24. Chen D, Bhambhvani HP, Hom J, et al. Effect of electronic clinical decision support on imaging for the evaluation of acute low back pain in the ambulatory care setting. World Neurosurg. 2020;134:e874–e877. INCOMPLETE (missing DOI)
25. Zafar HM, Ip IK, Mills AM, Raja AS, Langlotz CP, Khorasani R. Effect of clinical decision support-generated report cards versus real-time alerts on primary care provider guideline adherence for low back pain outpatient lumbar spine MRI orders. AJR Am J Roentgenol. 2019;212(2):386–394. doi:10.2214/AJR.18.19780
26. Benditz A, Faber F, Wenk G, et al. The role of a decision support system in back pain diagnoses: a pilot study. Biomed Res Int. 2019;2019:1314028. doi:10.1155/2019/1314028
27. Benditz A, Pulido LC, Grifka J, Ripke F, Jansen P. A clinical decision support system in back pain helps to find the diagnosis: a prospective correlation study. Arch Orthop Trauma Surg. 2023;143(2):621–625. doi:10.1007/s00402-021-04080-y
28. Maheronnaghsh R, Nezareh S, Sayyah MK, Rahimi-Movaghar V. Developing SNOMED-CT for decision making and data gathering: a software prototype for low back pain. Acta Med Iran. 2013;51(8):548–553.
29. Fakharian E, Nabovati E, Farzandipour M, Akbari H, Saeedi S. Diagnosis of mechanical low back pain using a fuzzy logic-based approach. Int J Intell Syst Appl Eng. 2021;9(3):116–120. doi:10.18201/ijisae.2021.239
30. Kobayashi H, Sekiguchi M, Yonemoto K, et al; DISTO project working group. Reference values of the Japanese Orthopaedic Association Back Pain Evaluation Questionnaire in patients with lumbar spinal stenosis and characteristics of deterioration of quality of life: lumbar spinal stenosis diagnosis support tool (DISTO project). J Orthop Sci. 2019;24(4):584–589. doi:10.1016/j.jos.2018.11.022
31. Al-Kasasbeh R, Korenevskiy N, Ionescu F, Alshamasin M, Smith AP, Alwadie A. Biotechnical measurement and software system for the prediction and diagnosis of osteochondrosis of the lumbar region based on acupuncture points with the use of fuzzy logic rules. Biomed Tech. 2013;58(1):51–65. doi:10.1515/bmt-2012-0081
32. Tun Firzara AM, Teo CH, Teh SY, et al. Evaluation of an electronic clinical decision support system (DeSSBack) to improve low back pain management: a pilot cluster randomized controlled trial. Fam Pract. 2023;40(5–6):742–752. doi:10.1093/fampra/cmad044

33. Gal N, Stoicu-Tivadar V, Andrei D, Nemeş DI, Nădăşan E. Computer assisted treatment prediction of low back pain pathologies. In: Stoicu-Tivadar L, et al., editors. Cross-border challenges in informatics with a focus on disease surveillance and utilising big-data. Amsterdam: IOS Press; 2014. p.47–51. doi:10.3233/978-1-61499-389-6-4
34. 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. doi:10.1016/j.ijmedinf.2017.11.010
35. Altun S, Alkan A. Lumbar spinal stenosis analysis with deep learning based decision support systems. GU J Sci. 2023;36(3):1200–1215. doi:10.35378/gujs.1116423
36. Gambo I, Mbada C, Aina S, et al. Implementing a decision support tool for low-back pain diagnosis and prediction based on the range of motions. Indones J Electr Eng Comput Sci. 2024;33(2):1302–1314. doi:10.11591/ijeecs.v33.i2.pp1302-1314
37. Zhang D, Du J, Shi J, et al. A fully automatic MRI-guided decision support system for lumbar disc herniation using machine learning. JOR Spine. 2024;7(2):e1342. doi:10.1002/jsp2.1342
38. Øverås CK, Nilsen TIL, Nicholl BI, et al. Multimorbidity and co-occurring musculoskeletal pain do not modify the effect of the SELFBACK app on low back pain-related disability. BMC Med. 2022;20:53. doi:10.1186/s12916-022-02237-z
39. Sandal LF, Øverås CK, Nordstoga AL, et al. A digital decision support system (SELFBACK) for improved self-management of low back pain: a pilot study with 6-week follow-up. Pilot Feasibility Stud. 2020;6:72. doi:10.1186/s40814-020-00604-2
40. Lin L, Hu PJH, Sheng ORL. A decision support system for lower back pain diagnosis: uncertainty management and clinical evaluations. Decis Support Syst. 2006;42(2):1152–1169. doi:10.1016/j.dss.2005.10.007
41. Kadhim MA. FNDSB: a fuzzy-neuro decision support system for back pain diagnosis. Cogn Syst Res. 2018;52:691–700. doi:10.1016/j.cogsys.2018.08.021
42. Badahman F, Alsobhi M, Alzahrani A, et al. Validating the accuracy of a patient-facing clinical decision support system in predicting lumbar disc herniation: diagnostic accuracy study. Diagnostics. 2024;14(17):1870. doi:10.3390/diagnostics14171870

Correspondence:
Received: 16 September 2025
Accepted: 21 October 2025

Corrado Ciatti, MD
Department of Orthopaedics and Traumatology, Guglielmo da Saliceto Hospital, Via Taverna 49, 29121, Piacenza (PC), Italy; University of Parma
E-mail: dadociatti@icloud.com
ORCID: 0000-0002-7094-4344

Appendix – Supplementary Material A – Search Strategy

Final search date: 14 July 2025. Databases: PubMed, Scopus, ProQuest, PsycINFO. No language or date limits were applied at the search stage.

Table A1. Records identified per database

Database	Records identified (n)
PubMed	287
Scopus	1715
ProQuest	250
PsycINFO	48

A.1 PubMed (MEDLINE via PubMed)

Query (copy/paste):

("Low Back Pain"[Mesh] OR "low back pain"[tiab] OR lumbago[tiab] OR "lumbar radiculopathy"[tiab] OR sciatica[tiab]) AND ("Clinical Decision Support Systems"[Mesh] OR "Decision Support Systems, Clinical"[Mesh] OR "clinical decision support" [tiab] OR "decision support system*"[tiab] OR cdss[tiab] OR "decision aid*"[tiab] OR "expert system*"[tiab] OR "rule-based"[tiab] OR "knowledge-based"[tiab] OR "computerized decision"[tiab]

OR “machine learning”[tiab] OR “artificial intelligence”[tiab])

Notes: Searched in All Fields with MeSH explosion. No study design or language filters applied at search.

A.2 Scopus (Elsevier)

Query (copy/paste):

TITLE-ABS-KEY ((“low back pain” OR lumbago OR “lumbar radiculopathy” OR sciatica) AND (“clinical decision support” OR “decision support system*” OR cdss OR “decision aid*” OR “expert system*” OR “rule-based” OR “knowledge-based” OR “computerized decision” OR “machine learning” OR “artificial intelligence”))

Notes: Documents of all types; no language limits at search.

A.3 ProQuest (Health & Medicine)

Query (copy/paste):

TI,AB((“low back pain” OR lumbago OR “lumbar radiculopathy” OR sciatica) AND (“clinical decision support” OR “decision support system*” OR cdss OR “decision aid*” OR “expert system*” OR “rule-based” OR “knowledge-based” OR “machine learning” OR “artificial intelligence”))

A.4 PsycINFO (via ProQuest)

Query (copy/paste):

((DE “Low Back Pain” OR TI,AB(“low back pain” OR lumbago OR “lumbar radiculopathy” OR sciatica)) AND (DE “Decision Support Systems” OR TI,AB(“clinical decision support” OR “decision support system*” OR cdss OR “decision aid*” OR “expert system*” OR “rule-based” OR “knowledge-based” OR “machine learning” OR “artificial intelligence”)))

A.5 Deduplication and screening

Results from all databases were exported in RIS/CSV format and merged in a reference manager. Duplicates were removed using exact matches on DOI/PMID/title/author (duplicates removed: n = 374). After deduplication, records screened: n = 1,926. Title/abstract screening and full-text assessment were conducted in pairs.

Supplementary Material B – Risk of Bias

We evaluated risk of bias using design-appropriate tools: QUADAS-2 (with AI-specific items) for diagnostic-accuracy and model development/validation studies; RoB 2 for randomized trials; and ROBINS-I for non-randomized comparative studies. Full criteria and per-study judgments are tabulated below.

Study (Author, Year)	RoB: Patient selection	RoB: Index test	RoB: Reference standard	RoB: Flow/ timing	Overall RoB
Lin Lin et al. (2006)	Some concerns	Low	High	Low	High
Kobayashi et al. (2019)	Low	Low	High	Low	High
Oude Nijeweme-d'Hollosy et al. (2018)	High	Low	Some concerns	Low	High
Kadhim et al. (2018)	High	Some concerns	High	Some concerns	High
Al-Kasasbeh et al. (2013)	High	Some concerns	High	Some concerns	High
Altun & Alkan (2023)	Some concerns	Low	Some concerns	Low	Some concerns
Gambo et al. (2024)	Some concerns	Low	Unclear	Low	Some concerns
Benditz et al. (2019)	Some concerns	Low	High	Low	High
Badahman et al. (2024)	High	High	Some concerns	Low	High
Fakharian et al. (2021)	High	Some concerns	High	Some concerns	High
Zhang et al. (2024)	Some concerns	Low	Low	Low	Some concerns
Hamtaei Pour Shirazi et al. (2025)	Some concerns	Some concerns	High	Low	High
Gal et al. (2014)	High	Some concerns	High	Some concerns	High

Figure S1A. QUADAS-2 risk of bias across studies (traffic-light).

Trial (Author, Year)	Randomization process	Deviations from intended interventions	Missing outcome data	Measurement of the outcome	Selection of the reported result	Overall bias
Tun Firzara Abdul Malik et al. (2023)	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns
Benditz et al. (2023)	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns
Øverås et al. (2022)	Low	Some concerns	Low	Some concerns	Low	Some concerns
Zafar et al. (2019)	Low	Low	Low	Low	Low	Low

Figure S1B. RoB 2 (randomized trials) traffic-light.

Study (Author, Year)	Confounding	Selection of participants	Classification of interventions	Deviations from intended interventions	Missing data	Measurement of outcomes	Selection of reported result	Overall bias
Chen et al. (2020)	Serious	Low	Low	Low	Low	Low	Some concerns	Serious
Benditz et al. (2023)	Serious	Moderate	Low	Low	Moderate	Moderate	Some concerns	Serious
Sandal et al. (2020)	Serious	Moderate	Low	Low	Moderate	Moderate	Some concerns	Serious

Figure S1C. ROBINS-I (non-randomized studies) traffic-light.

Table S2A. QUADAS-2 criteria (with AI-specific items)

Domain	Signalling question	Response options	Judgement rule
Patient selection	Consecutive/random sample enrolled? Case–control avoided? Inappropriate exclusions avoided?	Yes/No/Unclear	Low if Yes; High if case–control/non-consecutive/inappropriate exclusions; Unclear if insufficient detail
Index test (CDSS)	Index test interpreted without knowledge of reference standard? Threshold prespecified?	Yes/No/Unclear	Low if blinded/automated and prespecified; High if not; Unclear if not reported
Reference standard	Reference standard likely classifies target condition correctly? Interpreted without knowledge of index test?	Yes/No/Unclear	Low if appropriate and blinded; High if inappropriate; Unclear if not reported
Flow & timing	Appropriate interval between index and reference? Same reference for all? All included in analysis?	Yes/No/Unclear	Low if acceptable and complete; High if differential verification/exclusions; Unclear if not reported
AI-specific	Data leakage avoided (independent train/valid/test; no tuning on test)?	Yes/No/Unclear	Low if clearly avoided; High if probable; Unclear if not described
AI-specific	Class imbalance/missing data appropriately handled?	Yes/No/Unclear	Low if handled (weighting/stratification/imputation); High if ignored; Unclear if not reported
AI-specific	External validation performed?	Yes/No/Unclear	Low if independent dataset; High if only internal; Unclear if not reported

Table S2B. RoB 2 criteria

Domain	Signalling question	Judgement options
Randomization process	Random allocation & concealment; baseline balance assessed	Low/Some concerns/High
Deviations from intended interventions	Blinding and adherence; analytic approach appropriate	Low/Some concerns/High
Missing outcome data	Low loss; reasons unrelated to outcome; appropriate handling	Low/Some concerns/High
Measurement of the outcome	Valid/consistent measures; blinded assessors	Low/Some concerns/High
Selection of the reported result	Pre-specified analysis; no selective reporting	Low/Some concerns/High

Table S2C. ROBINS-I criteria.

Domain	Signalling question	Judgement options
Confounding	Important confounders measured/adjusted	Low/Moderate/Serious/Critical
Selection of participants	Appropriate inclusion into cohorts	Low/Moderate/Serious/Critical
Classification of interventions	Correct, consistent classification	Low/Moderate/Serious/Critical
Deviations from intended interventions	Deviations unlikely to bias	Low/Moderate/Serious/Critical
Missing data	Low missingness or appropriate handling	Low/Moderate/Serious/Critical
Measurement of outcomes	Outcome measurement appropriate/blinded	Low/Moderate/Serious/Critical
Selection of reported result	No selective reporting	Low/Moderate/Serious/Critical

Table S3A. QUADAS-2 per-study judgments and applicability

Study (Author, Year)	Design	RoB: Patient selection	RoB: Index test	RoB: Reference standard	RoB: Flow/ timing	AI: Data leakage	AI: Threshold prespecified	AI: Class imbalance/ missing	AI: External validation	Applicability: Patient selection	Applicability: Index test	Applicability: Reference standard	Overall RoB	Notes/ Justification
Lin Lin et al. (2006)	Knowledge base verification, Turing test, clinical efficacy study	Some concerns	Low	High	Low		Yes		Unclear	Some concerns	Low	High	High	Rule-based verbal-probabilities; evaluation vs clinicians; clinician diagnosis as reference
Kobayashi et al. (2019)	Multicenter cross-sectional study	Low	Low	High	Low		Yes		Unclear	Low	Low	High	High	Multicentre cross-sectional; DST used to classify LSS; outcome = JOABPEQ; no independent gold standard
Oude Nijeweme-d'Hollosy et al. (2018)	Model development and validation (ML)	High	Low	Some concerns	Low	Unclear	Unclear	Unclear	Unclear	Some concerns	Low	Some concerns	High	Training on vignettes; tested on small real-life set; reference = clinician referral
Kadhim et al. (2018)	System development, pilot evaluation	High	Some concerns	High	Some concerns	Unclear	Unclear	Unclear	No	High	Some concerns	High	High	n=10 case study; fuzzy-neuro; internal evaluation only
Al-Kasasbeh et al. (2013)	System development, expert-driven validation	High	Some concerns	High	Some concerns	Unclear	Unclear	Unclear	No	High	High	High	High	Acupuncture point energy; fuzzy + certainty factors; engineering study; limited clinical standard
Altun & Alkan (2023)	Model development, performance comparison	Some concerns	Low	Some concerns	Low	Unclear	Unclear	Unclear	No	Some concerns	Low	Some concerns	Some concerns	Deep-learning segmentation; dataset-based; no clinical threshold outcome
Gambo et al. (2024)	Model development, simulation/ validation	Some concerns	Low	Unclear	Low	Unclear	Unclear	Unclear	No	Some concerns	Low	Unclear	Some concerns	Open-source dataset; ensemble ML; no external validation; referral types based on dataset labels

Benditz et al. (2019)	Correlational pilot (DSS vs. surgeon)	Some concerns	Low	High	Low		Yes		Unclear	Some concerns	Low	High	High	Outpatient clinic; surgeon diagnosis as reference; physicians blinded to DSS
Badahman et al. (2024)	Diagnostic accuracy (Delphi, MRI as gold standard)	High	High	Some concerns	Low	Unclear	Unclear	Unclear	Unclear	High	High	Some concerns	High	Small convenience series; thresholds not prespecified; MRI reference
Fakharian et al. (2021)	ANFIS model, diagnostic accuracy	High	Some concerns	High	Some concerns	Unclear	Unclear	Unclear	No	High	Some concerns	High	High	Retrospective; internal split only; labels from physician record
Zhang et al. (2024)	Dev/ validation/ crossover (AI vs. surgeons)	Some concerns	Low	Low	Low	Unclear	Unclear	Unclear	No	Some concerns	Low	Low	Some concerns	Single-centre MRI; ground truth by radiologists; internal train/test
Hamtaei Pour Shirazi et al. (2025)	System development, diagnostic accuracy, SVM	Some concerns	Some concerns	High	Low	Unclear	Unclear	Unclear	No	Some concerns	High	High	High	iEMG/ biotechnical signals; internal CV; expert-derived labels
Gal et al. (2014)	System development, initial test	High	Some concerns	High	Some concerns	Unclear	Unclear	Unclear	No	High	Some concerns	High	High	Zebris mobility device; fuzzy rules; outcome = treatment recommendation; limited clinical validation

Table S3B. RoB 2 per randomized trial

Trial (Author, Year)	Design	Randomization process	Deviations from intended interventions	Missing outcome data	Measurement of the outcome	Selection of the reported result	Overall bias	Notes
Tun Firzara Abdul Malik et al. (2023)	Pilot cluster RCT (2 months)	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	Pilot cluster RCT; small clusters; limited blinding
Benditz et al. (2023)	Prospective, nonrandomized, unblinded	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	Preliminary judgement; to be verified in full text
Øverås et al. (2022)	RCT, 9m follow-up, secondary analysis	Low	Some concerns	Low	Some concerns	Low	Some concerns	Block randomization; not blinded; self-reported outcomes
Zafar et al. (2019)	Cluster RCT, 3 arms, multi-period	Low	Low	Low	Low	Low	Low	Practice-level randomization; objective EHR imaging-order outcomes

Table S3C. ROBINS-I per non-randomized study

Study (Author, Year)	Design	Confounding	Selection of participants	Classification of interventions	Deviations from intended interventions	Missing data	Measurement of outcomes	Selection of reported result	Overall bias	Notes
Chen et al. (2020)	Before/after implementation study	Serious	Low	Low	Low	Low	Low	Some concerns	Serious	Pre–post implementation without concurrent control; objective EHR outcomes; likely secular trends
Benditz et al. (2023)	Prospective, nonrandomized, unblinded	Serious	Moderate	Low	Low	Moderate	Moderate	Some concerns	Serious	Nonrandomized design; no concurrent control; small sample
Sandal et al. (2020)	Single-arm pilot, 6wk FU	Serious	Moderate	Low	Low	Moderate	Moderate	Some concerns	Serious	Nonrandomized design; no concurrent control; small sample

Table S4. Full-text reports excluded and reasons (n = 34)

Reason for exclusion at full-text	n
Not peer-reviewed sources	2
Review articles	9
Outcome not of interest/invalid	10
Protocol/design papers	8
Retracted/withdrawn	2
Not a CDSS	3

Two reports were not retrieved (n = 2).