Invasive Physiotherapy and Musculoskeletal Medicine

eISSN: 3101-0105 invasivephysiotherapy.com

doi: 10.63360/ipmm.v2.e25

Scoping Review

Use of Artificial Intelligence as a Clinical Decision Support System in the Management of Chronic Low Back Pain. A Scoping Review with Applied Conceptual Component

Marcos Campillos Lázaro[1] 0009-0001-6347-7290

[1] Faculty of Health Sciences, University of Zaragoza, Zaragoza, Spain.

ABSTRACT

Background and Objectives: The high prevalence of chronic low back pain and the growing demand for care require more efficient physiotherapy management models. Artificial Intelligence (AI) is presented as a support tool, although its clinical integration raises ethical and technical questions. The main objective was to review the recent literature on clinical decision support systems (CDSS) in lumbar physiotherapy and to explain their implications for the design of supervised clinical assistants.

Methods: A scoping review with a conceptual component was conducted using structured searches in PubMed, Web of Science, Scopus, Cochrane, and PEDro. Original studies published in the last 5 years that applied machine learning algorithms or digital tools in the management of low back pain were selected.

Results: Evidence indicates that machine learning algorithms have a high predictive capacity for triage and functional prognosis, showing equivalent or sometimes superior performance to conventional tools. In the therapeutic field, digital interventions show positive results for adherence and remote monitoring, although their direct efficacy on pain varies. The lack of transparency of algorithms was identified as the main barrier to trust.

Discussion: The results suggest that the current usefulness of AI lies in safety filtering and patient stratification, rather than in autonomous prescribing. The need to implement models with mandatory human supervision to mitigate risks is discussed. As a practical implication of the review, the theoretical feasibility of assistants based on strict clinical rules for detecting red flags and on RAG (Retrieval-Augmented Generation) architectures to personalize education without hallucinations is proposed.

Conclusions: AI is a valid technology for supporting triage and optimizing resources, but it requires external validation and transparency. The profession must lead the integration of these digital "co-pilots" to offload bureaucratic tasks and strengthen the therapeutic alliance, avoiding unsupervised automation.

Keywords: Artificial Intelligence; Clinical Decision Support Systems; Physiotherapy; chronic low back pain; Machine Learning; Telerehabilitation

4.0 BY-NC-SA

Correspondence: 869739@unizar.es

1. Introduction

Low back pain (LBP) persists as one of the main causes of years lived with disability globally, a public health challenge whose prevalence has increased by 60% in the last three decades[1,2]. In its chronic form (CLBP), defined as having an evolution of more than three months, it represents the subgroup with the greatest clinical, economic, and functional impact and constitutes the central focus of this work. This burden is not limited to individual suffering or functional restriction, but generates considerable economic pressure on health systems[3]. In the United States, direct costs exceed $134.5 billion[1]. In Spain, LBP has a prevalence of 17.1% in men and 24.5% in women, with a total attributable cost of almost 9,000 million euros (0.68% of GDP)[4]. This entails a great personal burden, a very high use of healthcare resources, and a significant loss of work productivity, a problem that extends to the continental level, where 43% of European workers report experiencing back pain[5]. Despite this massive investment, disability indicators continue to rise, suggesting that current health management models require a thorough review against the criteria of efficiency and equity[1,6,7].

The complexity of the approach to chronic low back pain lies in the need to articulate a biopsychosocial framework that goes beyond the biomedical paradigm and accommodates patient heterogeneity[1,6,7]. However, in daily practice, the physiotherapist faces information overload[3]. The constant updating of the evidence is unmanageable, and there are difficulties in objectively quantifying the evolution of pain or monitoring patient adherence outside the clinic[2,8]. In this context of saturation and clinical variability, Artificial Intelligence (AI), and specifically Machine Learning (ML), emerges as a tool capable of identifying nonlinear patterns in large volumes of data without the need for explicit manual programming[1,9,10].

The application of these technologies usually manifests in Clinical Decision Support Systems (CDSS), designed not to replace professional judgment but to organize multimodal information, such as medical history, images, and psychosocial variables, and to support diagnostic or therapeutic decisions[3,11,12]. However, the integration of AI into physiotherapy poses ethical and operational challenges, especially regarding the supervision of models that, although promising, lack deep clinical reasoning[9,13,14].

Given the nascent nature of this technology and the great heterogeneity of the tools available (ranging from predictive machine learning algorithms to conversational systems based on Large Language Models (LLMs))[1,13], the current literature is highly dispersed and makes it difficult to synthesize results and draw firm conclusions using strict quantitative methods such as meta-analysis[15]. For this reason, the present study adopts the design of a scoping review, a typology that allows for the systematic mapping of the literature on an emerging and heterogeneous field without being restricted to a single clinical question. This approach is ideal for identifying trends, critically evaluating the real clinical utility of these tools, and, on that basis, proposing applied conceptual frameworks, rather than being limited to a premature statistical aggregation of data.

The main objective of this work is to analyze recent scientific evidence on the applications of AI and CDSS systems in the management of chronic low back pain. Beyond making a descriptive synthesis of the literature, the central purpose is to use this evidence as a basis for the theoret-

ical design of a totally new conceptual framework of "clinical co-pilot", aimed at optimizing triage, assessment, and exercise prescription, and that guarantees patient safety through human supervision (Human-in-the-loop) in the face of automation.

2. Methods

A scoping review has been carried out with an applied conceptual component, with a systematized search strategy, aimed at integrating heterogeneous evidence on an emerging technological topic. Although the synthesis of the results is narrative, principles of transparency and reproducibility, adapted from the PRISMA 2020 declaration[16], were applied to the study identification and selection process.

2.1 Protocol and Registration

The present work does not have a protocol prospectively registered on platforms such as PROSPERO or OSF, given its nature as a scoping review in an emerging field. However, the methodology follows a structured and replicable workflow described below, specifically designed to map the recent literature and support the proposal for technological intervention, without claiming the statistical rigor of a meta-analysis.

2.2 Eligibility Criteria

The selection criteria were defined under the PICO framework (Population, Intervention, Comparison, Outcomes):

- Population: Adult patients (>18 years old) diagnosed with chronic low back pain, clinically defined as having an evolution of more than 3 months. Exceptionally, studies involving patients in the acute phase were included only when the AI tool's primary objective was to predict the risk of long-term pain chronicity. Similarly, studies on chronic pain that were not specifically low back pain were included when their main objective was phenotyping or stratifying patients using AI algorithms, given the direct applicability of their clustering methodologies to the management of chronic low back pain.
- Intervention: Implementation of Clinical Decision Support Systems (CDSS) based on Artificial Intelligence (AI), Machine Learning algorithms, or Language Models (LLMs).
- Comparison: Studies comparing the intervention with conventional physiotherapy were included, as well as technical validation studies without a clinical control group.
- Outcomes: We included studies evaluating the effectiveness of AI tools for triage, assessment, and exercise prescription. Specifically, a heterogeneous spectrum of variables grouped into three categories was analyzed: 1. Technical and predictive accuracy variables (performance of algorithms for triage and prognostic prediction); 2. Clinical outcome variables linked to assessment and treatment (pain intensity, functional disability, quality of life, and psychosocial factors); 3. Operational and process variables (adherence to the exercise prescription, usability of the system, and feasibility of implementation).

Regarding the rest of the inclusion criteria, randomized clinical trials (RCTs), cohort studies (prospective and retrospective), pilot studies, and validation work on predictive models that provid-

ed empirical data were admitted. We also limited the search to English- or Spanish-language articles published in the last 5 years (2020-2025), excluding single-case reports, opinion articles without data, and oncology medical imaging studies unrelated to rehabilitation.

2.3 Data Sources and Search Strategy

The literature search was conducted in November 2025 across five international databases: PubMed, Scopus, Web of Science, PEDro, and Cochrane Library (see the detailed strategy in Table S1 of the Supplementary Material). A "Master Equation" was designed combining controlled terms (MeSH) and free language using Boolean operators, structured in three thematic blocks:

1. Technology: “Artificial Intelligence”, “Machine Learning”, “Deep Learning”, “CDSS”, “Generative AI”.
2. Intervention: “Physiotherapy”, “Physical Therapy”, “Rehabilitation”, “Telerehabilitation”.
3. Pathology: "Low Back Pain", "Chronic Low Back Pain". In addition, a reverse manual search was performed in the bibliographic references of the selected studies ("snowball").

2.4 Selection of Studies included in the Review

The selection process was carried out by a reviewer in two sequential phases to ensure methodological rigor. In the first stage, a screening of the identified records was conducted by reading the titles and abstracts, discarding those that did not fit the study's subject. In the second stage, the full texts of the shortlisted articles were examined to verify strict compliance with the eligibility criteria described above. Final selection and data extraction were supervised by a second independent reviewer to verify the consistent application of the eligibility criteria, ensure traceability of the process, and confirm that the studies met the inclusion criteria. This reviewer is not listed as a co-author of the manuscript because he or she does not meet the authorship criteria established by the ICMJE.

It should be noted that, consistent with the design of a scoping review with a structured search, the initial screening was not carried out independently in duplicate, so the second reviewer's role was limited to supervising the process and not to calculating interobserver agreement.

2.5 Data Extraction Process

The relevant information from each study was extracted and organized in a synthesis table (Table 1) considering the following variables: authorship/year, methodological design, sociodemographic characteristics of the sample (size, participant profile, occupation, and other relevant details), AI tool used (e.g., computer vision, machine learning), clinical objective (diagnosis, prognosis or treatment) and main findings.

For the management and organization of the files of the included studies, as well as for the grammatical and stylistic review of the draft, assistance tools based on Generative AI (Google Gemini and Google NotebookLM) were used. In no case were these tools delegated the critical analysis of the evidence, the interpretation of the results, or the decision-making on the structure or content of the manuscript, responsibilities that fall exclusively on the authors. All content was checked by the authors against the original sources before inclusion in the paper.

2.6 Results Synthesis Process

Given the clinical and technological heterogeneity of the studies (ranging from algorithm validations to telerehabilitation trials), a meta-analysis was not feasible. A thematic narrative synthesis was chosen, grouping the results into three main categories: diagnostic and prognostic accuracy of the models, therapeutic effectiveness of AI-supported interventions, and process variables related to adherence and usability of digital tools. This approach enabled the integration of quantitative and qualitative evidence to provide a holistic view of the state of the art.

2.7 Study Variables

The variables of interest analyzed included:

- Diagnostic and assessment capacity: accuracy of the models to identify pathologies, classify pain phenotypes, and perform initial triage (measured by area under the curve (AUC) and other predictive performance metrics).
- Prognostic accuracy: AI's ability to anticipate the patient's evolution and response to treatment (risk of chronification, probability of functional improvement).
- Treatment effectiveness and adequacy: impact of proposed or AI-guided interventions in terms of clinical improvements, including pain intensity (Visual Analogue Scale, VAS / Numeric Rating Scale (NRS), functional disability (Oswestry Disability Index, ODI / Roland Morris Disability Questionnaire (RMDQ), quality of life and, where available, relevant physical and psychosocial variables (e.g. motor strength or control, anxiety, depression or catastrophizing of pain).
- Process and engagement variables: level of adequacy of the technological proposal in terms of patient adherence, program completion rates, usage patterns (frequency and duration of sessions), and usability/acceptability of digital tools

2.8 Assessment of Methodological Quality/Risk of Bias

The evaluation of methodological quality was adapted to the specific design of each research. For randomized clinical trials, methodological quality was assessed using the PEDro scale[17] (breakdown of scores in Table S2 of the Supplementary Material). For observational and cohort studies (retrospective and prospective), methodological quality was assessed using the Newcastle-Ottawa Scale (NOS)[18] (breakdown of scores in Table S3 of the Supplementary Material), which evaluates group selection, comparability, and outcome ascertainment.

3. Results

3.1 Selection of Studies

The identification and selection flow is illustrated in the PRISMA diagram (Figure 1). The initial search of the databases yielded a total of 661 records. After the elimination of 179 duplicates, 482 articles were screened. Of these, 401 were discarded after title review, and 81 abstracts were evaluated. Finally, after reading 40 full-text reports, 21 were excluded for inadequate design or lack of quantitative results, resulting in a final sample of 19 studies included in the review.

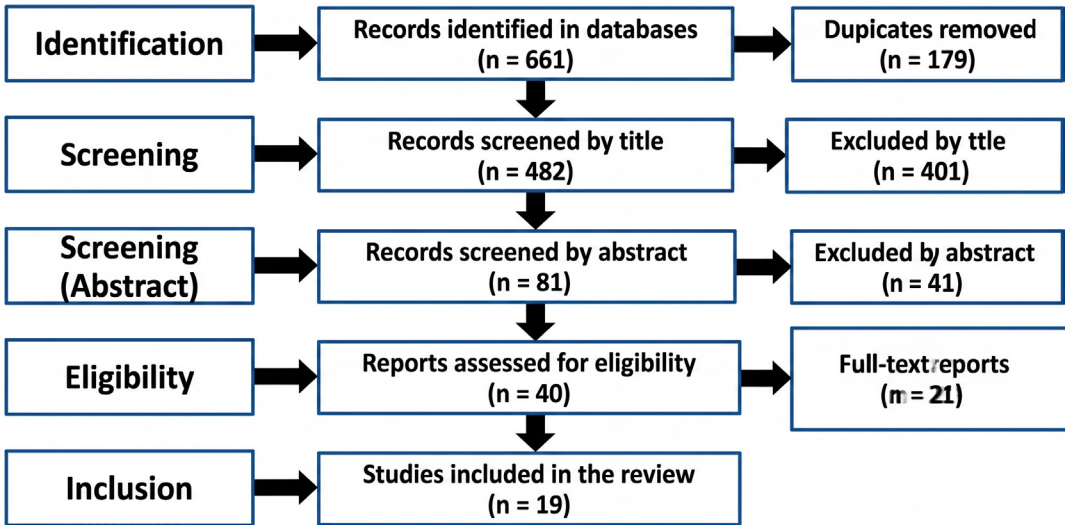

Figure 1. PRISMA 2020 Flowchart. Source: Authors' elaboration assisted by Generative AI tools (Google Gemini)

3.2 Characteristics of the Studies

The analysis covers nineteen original articles published between 2021 and 2025, which confirms the emerging nature of this line of research. The geographical distribution is heterogeneous, with research carried out mainly in Europe, the United States, and Asia. Regarding the methodological design, the evidence is grouped into two large blocks:

- Observational studies and validation of predictive models (n=12) focused on the processing of massive data using machine learning algorithms.
- Randomized clinical trials (n=7) aimed at comparing the therapeutic efficacy of AI-designed or monitored digital interventions versus standard care.

Detailed information on the design, intervention, and results of each study is presented in Table 1.

Table 1: Summary of the characteristics and main findings of the included studies (2020-2025).

Author and Year	Design, Sample & Profile	AI Tool	Comparator	Clinical Objective	Main results	Other results
Xiao et al., 2025 [19]	RCTs (n=38 randomized, 34 completed). Mean age 29 years, 76.5% women. High schooling (median 16-17 years of education) and normal BMI (~21 kg/m²). Chronic non-specific low back pain	Computer Vision. The Yun-fu app uses AI to identify 30 key body points using the camera, recognizing movements and providing real-time guidance and feedback during exercises	The group that received conventional video guidance	Treatment (Application), as AI is used to give real-time guidance and feedback during exercise	Patients treated with AI showed significant improvements compared to control in the reduction of maximum (p<0.001) and mean (p=0.01), functional disability (p=0.002), catastrophic thinking (p=0.04), and core muscle thickness (p<0.05) at four weeks	There were no significant differences between the groups regarding current pain (p=0.83), severe disability (p=0.07), and physical mobility (p=0.55) after four weeks of intervention
Cheema et al., 2025 [20]	Retrospective observational cohort study (n=2497 analyzed). Mean age 57 years. BMI > 30 present in 42.1% - 44.5% of patients	Machine Learning. The Random Forest algorithm analyzes 47 clinical variables of patients to predict the prognostic factors that determine treatment outcomes for chronic low back pain	N/A (Predictive Study)	Decision support (Prognosis), as AI is used to identify prognostic variables of outcomes in chronic low back pain	The predictive artificial intelligence model achieved 98.0% accuracy in identifying patients with greater improvements in functional status, outperforming linear regression models. The main prognostic variables included expected functional change, expected visits, age, and educational level of the practitioner	There were no significant differences in the clinical improvement of patients according to the specialized training of the physiotherapist (p=0.7546) or the history of previous surgeries (p=0.6218) using traditional statistical methods

Author and Year	Design, Sample & Profile	AI Tool	Comparator	Clinical Objective	Main results	Other results
Griefahn et al., 2025 [21]	Retrospective cohort analysis (n=379). Mean age was 51 years, 63.6% women. Occupation: Mostly sedentary or mixed (46.7% work sitting and 43.5% alternating sitting and standing). Patients with spinal pain (lumbar, cervical, dorsal, and buttocks)	Recommendation System. The medicalmotion app uses AI to dynamically select personalized exercises by analyzing the location and intensity of pain, medical history, and real-time patient feedback	N/A (Retrospective Cohort Study Without Control Group)	Treatment (Design), as AI is used to compose and personalize exercise programs for each patient	Patients treated with AI-compound exercises showed, without a control group, a significant reduction in pain intensity (p<0.001) and a substantial improvement in well-being (p<0.001) at eight weeks of intervention compared to their initial assessment	No statistically significant correlation was found between skipped exercises and pain intensity. As a secondary finding, the mean sessions lasted approximately ten minutes
Alzouhayli et al., 2025 [22]	Longitudinal sub-analysis of RCTs (n=52). Mean age was 44 years, 51.9% women. BMI: The mean BMI was 28.5 kg/m² (overweight). Patients with acute or recurrent non-specific low back pain	Recommendation System. The Tonal device uses AI to measure range of motion, power, and speed, providing the patient with real-time postural feedback and adjusting the resistance of the exercises	The group that received AI-guided therapy in the clinic	Treatment (Application), as AI is used to give real-time feedback and adjust resistance	Patients treated with AI-guided therapy at home showed clinically equivalent results to those in the clinical setting. Although statistically they showed less interference with pain, less kinesiophobia (p<0.01), and greater physical function (p<0.001), the improvement at one year was similar and without clinical relevance between groups	No significant differences were observed in the general clinical change from the end of the intervention to week 52 (p≥0.23), nor in the age of the participants (p=0.93)
Pfeifer et al., 2025 [23]	Cohort study (reanalysis of individual data from 2 RCTs; n=1,511). Mean age was 41 years, 57% female. BMI and Occupation: They had a mean BMI of 24.8 kg/m² (normal weight bordering on slight overweight) and the vast majority performed "white collar" or office jobs (n=1,402 vs. n=109 "blue collar"). Patients with non-acute non-specific low back pain	Machine Learning. Gradient Boosting Machines and Random Forests analyze clinical and psychosocial factors of the patient to predict their level of adherence to exercise programs for low back pain	N/A (Predictive Study)	Management variables (Adherence), as AI is used to identify predictors of exercise adherence	The predictive model achieved an accuracy of 64% to 83% in classifying treatment adherence. Major prognostic factors included fluctuations in pain intensity, baseline postural control, and reduced patients' exercise level	No significant differences were observed in adherence with respect to age (p=0.31), sex (p=0.26), or educational level (p=0.81). The overall mean adherence was 64%
Areias et al. (Scaling) 2024 [24]	Nonrandomized controlled study (post-hoc analysis of prospective cohorts; n=36,186 initiated, 27,741 completed). ~49 years old, ~58% female. High schooling (64-66% with university or higher education) and overweight (mean BMI ~29.5 kg/m²). Chronic musculoskeletal pain	Machine Learning. A machine learning system analyzes the patient's performance, pain, and fatigue to recommend adjustments to the physical therapist in the program, such as new exercises or dose changes	The standard digital care group is manually operated	Management variables (Scalability), as AI is used to recommend therapeutic adjustments and allow each physiotherapist to see more patients	Patients treated with the AI-backed program showed results equivalent to those of the standard control. No significant differences were observed in the rate of improvement response to pain (p=0.399), nor in the reduction of symptoms of anxiety (p=0.928) or depression (p=0.187)	The AI group achieved greater therapeutic adherence to the program (p<0.001). Specifically, there were no significant differences in the rate of adverse events (p=0.231) between the two groups
Areias et al. (Tool) 2024 [25]	Prognostic study (secondary analysis of prospective studies; n=6,125). 48.5 years, 56.2% women. High schooling (~66% with university or higher education) and overweight (mean BMI 29.5 kg/m²). Low back pain (81% chronic)	Machine Learning. RNN and LightGBM analyze the patient's movement, pain, and performance per session to predict their pain relief, alerting the physical therapist to adjust treatment	N/A (Predictive Study)	Decision support (Prognosis) uses AI to predict pain relief at the end of the program.	The predictive model showed progressively improving predictive capacity (p<0.001), achieving areas under the curve of 0.71 and 0.70 in the seventh session for identifying pain relief. The main prognostic variables included pain, physical performance, motivation, and adherence to treatment	No significant differences were found in predictive performance between neural network and decision tree models (p≥0.22). Secondarily, low adherence predicted a higher probability of failure

Author and Year	Design, Sample & Profile	AI Tool	Comparator	Clinical Objective	Main results	Other results
Zhao et al., 2024 [26] *	Registry-based cohort study (n=11,995). 43.9 years, 77.6% women. Predominantly secondary (~58%) or tertiary (~29%) schooling. Complex non-malignant chronic pain	Machine Learning. The LVMM algorithm classifies patients with chronic pain into four subgroups by analyzing clinical and psychosocial questionnaires to predict pain interference after one year of treatment	N/A (Predictive Study)	Decision support (Phenotyping), as AI is used to identify prognostic subgroups in patients with chronic pain	The predictive model explained 26% of the variance in pain interference at one year. He managed to classify patients into four distinct prognostic subgroups, showing statistically significant differences in general health status among all identified classes (p<0.001). The four subgroups were: Class 1 (21.5%): low pain but high emotional load; Class 2 (29.8%): greater clinical and psychological severity; Class 3 (30.3%): worse physical health with more moderate emotions; Class 4 (18.4%): healthier subgroup, less pain and better general condition.	No significant differences in pain interference were observed at one year of follow-up between the subgroup with low physical pain but high emotional burden (Class 1) and the healthier subgroup (p=0.130)
Bates et al., 2023 [27]	RCTs (n=108 randomized, 69 completed). 41.2 years, 65.2% women. Overweight (mean BMI 29.4 kg/m²). Low back pain (excluding patients with recent back surgery)	Machine Learning. Tonal AI analyzes the patient's power, range of motion, and speed in real-time to automatically adjust the training resistance and provide feedback on their technique	Control group that received no treatment	Treatment (Application), as AI is used to automatically adjust resistance during exercise	Patients treated with AI showed significant improvements compared to the control group in pain reduction (p≤0.05) and reduced interference with daily activities (p≤0.03). At eight weeks, this intervention achieved subjective clinical and functional outcomes equivalent to conventional therapeutic care.	The AI group did not achieve significant improvements in isometric extensor endurance (p≥0.17) or gait capacity (p≥0.39). All participants maintained an adherence rate of more than 66%.
Marcuzzi et al., 2023 [28]	3-arm RCTs (n=294 randomized). 50.6 years, 58.8% women. High schooling (61.2% with >12 years of education) and overweight (mean BMI 26.8 kg/m²). Patients with neck and/or low back pain are referred to specialty care	Case-Based Reasoning. SELFBACK uses case-based reasoning to analyze the patient's symptoms and daily steps, automatically generating personalized weekly exercise and education plans based on previous success stories	Two control groups: one group that received exclusive usual care (clinical examination and offer of appropriate treatment) and another group that received generic web support without personalization (e-Help) along with usual care	Management variables (Adherence), since AI is used to send interactive messages promoting the continuity of the exercises	Patients treated with the AI application showed results equivalent to those of usual care and web support. No significant differences were observed in the improvement of musculoskeletal health at three months compared to usual care (p=0.60) or web support (p=0.36)	There were no significant differences in disability, pain, or quality of life between groups. However, the AI group reported a significantly greater improvement in the patient's Overall Perceived Effect at 3 months compared to usual care (p=0.004). Secondarily, 29.3% of the AI group never accessed the app
Tun Firzara et al., 2023 [29]	RCT by pilot clusters (n=36 recruited). 47.4 years, 38.9% women. High schooling (41.7% have a university or bachelor's degree). Non-specific low back pain	Clinical Decision Support System. DeSSBack automatically analyzes a patient's nine-item questionnaire to stratify their risk level and recommend the appropriate treatment pathway	Group managed with regular clinical practice	Decision support (Triage), as AI is used to stratify patients' risk and guide their level of care	Patients treated using the electronic decision support system showed a greater reduction in low back disability than the usual control group, with a mean effect size of 0.718. Secondarily, small effect sizes were observed for the reduction of pain and depression, while the improvement in anxiety obtained a medium effect size (0.480).	Patients' therapeutic adherence to the prescribed treatment was poor. No significant baseline differences were observed between groups regarding pain intensity (p=0.318) or disability (p=0.548)
Park et al., 2023 [30]	RCTs (n=100 randomized and completed the study). 35.5 years, 40% women. Chronic low back pain	Deep Learning. Dr. AI analyzes the patient's movement through the mobile camera to diagnose, prescribe personalized exercises, and provide real-time audiovisual feedback that corrects their posture	Group that received conventional physical therapy	Treatment (Application), as AI is used to give real-time audiovisual feedback during exercises	Patients treated with AI showed results equivalent to those of conventional physiotherapy in reducing pain (p=0.41) and general disability (p=0.44). However, the AI intervention was superior in improving limitations in daily activities (RMDQ questionnaire, p=0.03), also showing to be cost-effective and safer	There were no significant differences between the groups in quality-of-life improvement (p=0.18). Secondarily, the AI group reported significantly greater accessibility (p=0.02), satisfaction with cost (p=0.01), and satisfaction with real-time feedback (p=0.01)

Author and Year	Design, Sample & Profile	AI Tool	Comparator	Clinical Objective	Main results	Other results
Hartmann et al., 2023 [31]	Longitudinal cohort study (n=58 recruited, 43 analyzed). 41.4 years (App group) and 50.9 years (Control group), 59% women (App group) and 50% women (Control group). Low back pain	Recommendation System. Medicalmotion uses AI analysis of anamnestic data, pain level, and user feedback to automatically recommend the most beneficial exercises from 350 options.	Usual care group with rehabilitation sport	Treatment (Design), as AI is used to choose and tailor exercise recommendations to the patient	Patients treated with AI showed significant improvements compared to the control group in pain reduction (p=0.003) and daily disability reduction (p<0.001) after eight weeks. However, pain reduction did not meet the threshold of clinical relevance (standardized to 2 points on the NRS scale)	As confounding factors, there were no significant differences in the evolution of pain with respect to age (p=0.23), sex (p=0.65), employment status (p=0.22), or sports activity (p=0.32)
Nordstoga et al., 2023 [32]	Secondary analysis of one RCT (n=461 randomized). ~47 years, 55.3% women. High schooling (between 58% and 72% of subgroups with >12 years of education) and overweight (mean BMI of ~27.6 kg/m²). Non-specific low back pain	Case-Based Reasoning. SelfBACK uses case-based reasoning to analyze patient characteristics, symptoms, and progression, tailoring personalized recommendations for physical activity, exercise, and education on a weekly basis	Group that received only usual care	Treatment (Design), as AI is used to tailor weekly self-management recommendations to the patient	The overall benefit of AI versus control was maintained regardless of pain duration or intensity. However, AI was significantly superior in reducing disability at three months, specifically in the subgroup with severe initial pain (>5/10), achieving an additional 1.8 points of improvement (p=0.03)	The superiority in disability observed at three months was not maintained at nine months. There were also no significant global differences in pain self-efficacy or perceived improvement
Strigo et al., 2023 [33] *	Retrospective cohort study (n=3,296 analyzed). 50–51 years, ~64.6% female. Obesity (mean BMI of the subgroups between 30.5 and 33.0 kg/m²). Patients with complex chronic pain refractory to routine interventions.	Machine Learning. Mclust analyzes nine initial symptoms reported by the patient using unsupervised clustering to classify them into three subgroups, predicting their trajectory and personalizing the treatment of chronic pain	N/A (Predictive Study)	Decision support (Phenotyping), as AI is used to identify phenotypes of chronic pain	The predictive model identified three clinical subgroups. All showed an overall clinical improvement (p=0.002) and a reduction in the impact of pain (p<0.001) after one year of treatment, although the most severe subgroup showed a smaller improvement at the end of follow-up (p<0.05)	There were no significant differences between subgroups in baseline blood pressure (p>0.05), proportion of patients medicated with opioids (p>0.1), or types of medical services used (p=1)
Zmudzki & Smeets, 2023 [34]	Retrospective cohort study (n=2,364 who completed the intervention). 43.8 years, 72.5% women. Predominantly secondary schooling (44.2% with vocational secondary education) and overweight or obesity (65.7% with BMI ≥ 25 kg/m²). Chronic musculoskeletal pain referred to an interdisciplinary multimodal rehabilitation program	Machine Learning. A machine learning framework analyzes 30 initial patient variables to predict their response to treatment in 13 dimensions, generating a prognostic profile that allows them to personalize their goals	N/A (Predictive Study)	Decision support (Prognosis), as AI is used to predict outcomes of multimodal treatment of chronic musculoskeletal pain	The predictive model achieved individual performance with AUCs between 0.49 and 0.65 across various clinical measures. However, by employing the full set of algorithms, the multidimensional profile consistently assessed and predicted treatment outcomes in 75.3% of patients.	Body mass index was not selected as a prognostic factor in any model. Universal predictive variables included treatment control, maximum pain, and employment status.
Knoop et al., 2022 [35]	Prospective cohort study (n=247 with baseline data). 49 years old, 41% women. Predominantly low schooling (60% with low educational level, corresponding to primary, secondary or vocational education). Acute low back pain (≤1 month duration)	Machine Learning. XGBoost analyzes three patient variables (resilience, prior disability, and recovery expectation) to automatically predict their risk of not recovering from low back pain at three months	N/A (Predictive Study)	Decision support (Prognosis), as AI is used to predict non-recovery from acute low back pain	The predictive model achieved an area under the curve of 0.66 and an accuracy of 63% for predicting lack of recovery from low back pain. Although it did not show superiority compared to the traditional statistical model (AUC 0.71), it exceeded the predictive capacity of the usual clinical assessment (AUC 0.53)	47% of participants did not recover after three months. Secondly, the inclusion of the clinical course of the first weeks did not substantially improve the performance of the model
Park et al., 2022 [36]	Retrospective case-control study (n=281 recruited, 176 analyzed). 36.7 years (App group) and 38.3 years (Control group), 60.6% women (App group) and 29.2% women (Control group). Office workers with low back pain	Deep Learning. MDMECA identifies 14 body points using the mobile camera to analyze movement in real time and provide visual and auditory feedback to the patient during exercise	The group that used YouTube exercise videos	Treatment (Application), as AI is used to analyze postures and give real-time feedback during exercise	Patients treated with the AI application showed greater improvements compared to video control in increasing quality of life (p<0.01) and achieved a greater reduction in the intensity of low back pain (p<0.01) after performing the exercise program	The group with AI registered greater adherence and satisfaction (p<0.01). There were no significant differences between the two groups regarding the economic willingness to pay for the program (p=0.59)

Author and Year	Design, Sample & Profile	AI Tool	Comparator	Clinical Objective	Main results	Other results
Anan et al., 2021 [37]	RCTs (n=121 randomized, 94 analyzed). ~42 years, 23.4% women. Office workers with neck/shoulder and/or lower back pain	Natural Language Processing. Secaide interacts daily through a mobile app to send short exercise routines, automatically adapting its responses and motivational messages according to the options selected by the patient	Regular Company Workout Routine	Management variables (Adherence), as AI is used to send interactive daily messages promoting the continuity of exercises	Patients treated with the AI-assisted program showed significant improvements compared to controls in reducing low back, neck, and shoulder pain at twelve weeks. In addition, they achieved a significantly higher rate of subjective clinical improvement of their symptoms (p<0.001)	As a secondary finding, the AI group achieved a high therapeutic adherence rate of 92%. In addition, daily exercise sessions required less than a minute to complete

Source: Authors' elaboration assisted by generative AI tools (NotebookLM)

*Study on complex chronic pain. Included for the methodological relevance of its phenotyping algorithm, applicable to the approach to chronic low back pain

3.3 Critical Evaluation

The methodological quality of the included randomized clinical trials was evaluated using the PEDro scale[17]. The mean score obtained was 5.7/10 (range 3-7) (see Table S2 for a detailed breakdown). The vast majority of studies (n=5) showed a high methodological quality (≥6/10) [19,28,29,30,32], standing out for the strict compliance with random assignment, intention-to-treat analysis, and rigorous statistical reporting. In contrast, two trials were at high risk of bias (≤4/10) [27,37]. The cross-sectional analysis reveals that the main methodological weakness in the current literature is the lack of blinding of participants and therapists, a difficulty inherent in interventions that combine physical exercise and technological devices[19,27,37]. However, this risk was partially mitigated by the successful implementation of blinding by independent evaluators in more than half of the trials. On the other hand, the methodological quality of the 12 observational studies (11 cohort and 1 case-control) was assessed using the Newcastle-Ottawa scale (NOS)[18], with a mean score of 6.4/9 (range 3-9) (see Table S3 for a detailed breakdown). Almost half of the sample (n=5) showed high quality (≥8/9)[23,24,25,26,34], demonstrating strong robustness to cohort representativeness and multivariate statistical adjustment. At an intermediate level, three studies showed a moderate risk of bias (6-7/9)[20,33,35]. However, one-third of the studies (n=4) were at high risk of bias (≤4/9)[21,22,31,36]. The main limiting factors in these lower-quality studies were a lack of comparability (not controlling for key confounders such as age or baseline severity) and deficiencies in follow-up, reporting significant losses that introduce potential attrition bias into the accumulated evidence.

3.4 Summary of Results

A distinctive feature of the analyzed literature is the multidimensionality of the data processed by Artificial Intelligence tools. Unlike traditional clinical models, the evaluated algorithms simultaneously integrated a wide spectrum of variables into their predictions and treatments. These can be classified into five major domains: 1) sociodemographic and anthropometric (e.g., age, sex, body mass index, employment status, and level of social deprivation); 2) clinical (e.g., pain intensity and duration, opioid use, and comorbidities); 3) functional and biomechanical (e.g., ODI and RMDQ indices, muscle strength, and range of motion captured by machine vision); 4) psychosocial (e.g., depression, anxiety, kinesiophobia, resilience, and recovery expectations); and 5) technological interaction, an emerging category that includes the frequency of use of ap-

plications, accuracy in the execution of exercises and real-time fatigue reporting. This richness of variables is what allows AI to perform much deeper phenotyping of patients than standard screening can.

3.4.1 Description of the results obtained for each variable

The findings were conceptually categorized by reported clinical utility: decision support (prognosis, phenotyping, and triage), treatment (design and application), and support and management variables (adherence and scalability).

Diagnostic and Prognostic Accuracy: Of the articles reviewed, 6 specifically analyze the use of machine learning algorithms for patient stratification and prognosis. Methodologically, these studies rely primarily on cohort designs (prospective and retrospective) and reanalysis of clinical databases to train algorithms. The evidence is mostly favorable: 5 of these studies suggest that machine learning offers greater predictive power for anticipating clinical outcomes[20,25,26,33,34]. However, 1 of the studies provides an important nuance by pointing out that, although machine learning outperforms standard screening questionnaires, its performance may be equivalent (and not necessarily superior) to that of traditional statistical regression models[35]. In general terms, machine learning algorithms showed, in most of the studies reviewed, a greater ability than conventional tools to stratify patients, although with variability between studies (such as the validated *STarT Back questionnaire*, the professional's own expectation, and clinical judgment):

- Random Forest algorithms identified functional improvement quartiles with 98% predictive accuracy for patients with better outcomes, outperforming classical linear regression models[20].
- The XGBoost predictive model showed better performance in predicting the risk of chronification (AUC 0.66; 95% CI: 0.56-0.70) compared with the *STarT Back questionnaire* (AUC 0.53; 95% CI: 0.45-0.60) in patients evaluated in the acute phase[35].
- Neural networks and LightGBM models were able to identify treatment-responding patients with an AUC of 0.71, enabling dynamic adjustments to the care plan[25].

In short, machine learning offers a predictive capacity superior to or equivalent to traditional tools to anticipate the risk of chronicity

Therapeutic Effectiveness (Pain and Disability): Of the articles reviewed, 8 directly evaluate the clinical efficacy of these digital interventions compared to control groups or face-to-face treatments[19,21,28,30,31,32,36,37]. It is worth noting the high level of evidence in this section, since 5 of these studies are randomized clinical trials[19,28,30,32,37], supported by 3 observational studies (cohorts and case-controls)[21,31,36]. The evidence reflects a mixed trend: 5 studies report favorable results of AI in reducing pain or disability[19,21,31,36,37], 1 study reports efficacy equivalent to traditional therapy[30], and 2 trials point to the absence of additional global benefits, conditioning their success to the initial severity of the patient[28,32]. Overall, the clinical efficacy of AI-based digital interventions is variable and inconsistent, underscoring the need to consider the patient's clinical profile when selecting and implementing these tools:

- Guided machine vision systems were associated with statistically significant reductions in pain compared with control groups ($p<0.001$), with reductions of up to 3.00 points on the NRS scale[19].
- Digital interventions significantly reduced lumbar functional disability, as evidenced by improvements of up to 3.8 points on the ODI index, and improved patients' quality of life[21,31]. In addition, they showed a clinical efficacy equivalent to face-to-face physiotherapy ($p=0.44$), standing out as a significantly more cost-effective alternative[22,30].
- However, some trials found no significant additional benefits in overall musculoskeletal health compared with usual care (reporting only a 0.62-point difference on the MSK-HQ questionnaire)[28], suggesting efficacy that depends on the patient's severity profile. For example, AI has been observed to achieve greater reductions in disability when applied to patients with a high baseline pain intensity (>5/10) compared to milder cases[32]
- However, in some studies, the magnitude of pain reduction, although statistically significant, did not meet the recommended minimum clinically important difference (MCID) threshold for chronic low back pain (defined as a reduction of ≥2 points on the NRS). For example, interventions have been reported to achieve absolute reductions of only 1.4 points, which limits the practical impact of these tools on the patient's perceived improvement[31].

In summary, the clinical efficacy of AI-based interventions is inconsistent; they show variable success that depends on the severity of the patient, and do not always reach clinical relevance

Moving from treatment-oriented studies to those focused on the operational viability of the system, a third block of evidence specifically analyzes the adherence, usability, and scalability of digital tools

Adherence, Usability, and Scalability: Of the articles reviewed, at least 5 specifically evaluate the feasibility of AI from the perspective of patient behavior and health system management. At the methodological level, this evidence is supported by a cross-sectional synthesis of randomized clinical trials, prospective cohort studies, and training of predictive models on clinical databases. Studies suggest the promising feasibility of these technologies in terms of safety and acceptability. However, heterogeneous results are observed in patient use, adherence, and engagement, which depend on the specific tool and the context of application:

- Interventions using *chatbots* resulted in 92% of patients assigned to this technology (44 out of 48 users) completing the 12-week exercise program[37].
- The Artificial Intelligence models were able to correctly predict in 83% of the cases evaluated (predictive accuracy) which patients would or would not comply with their face-to-face physiotherapy treatment (defining compliance as attendance at more than 70% of the sessions of a 3-week supervised protocol, at a rate of 3 weekly sessions), using variables such as postural control, pain fluctuations and the patient's previous level of physical activity[23].
- In terms of safety and scalability, AI-backed digital interventions showed a low risk profile, with low adverse event rates comparable to standard care (reporting incidences of only 0.58% in the AI-assisted group versus 0.69% in the control group; $p = 0.231$), without showing significant increases in undesirable effects associated with the use of technology[24].

In summary, digital tools show promising viability and a good security profile, although with heterogeneous results in use, adherence, and engagement depending on the tool and the context

3.4.2 Relationship between the results of the different variables

Cross-analysis reveals significant associations between clinical variables, adherence, and predictive efficacy:

Pain intensity and fluctuations: Some studies suggest that the effectiveness of digital support may depend on the initial severity. Patients with high pain (>5) achieved greater reduction in disability compared to control (p=0.03 interaction), while mild cases did not show additional benefits[32]. In addition, constant fluctuations in pain acted as the main predictor of low adherence[23].

In addition, the relationship between pain evolution and adherence is not linear but follows an inverted "U" shape. Predictive models suggest that patients tend to drop out of digital therapy, whether they experience marked increases in pain (for fear that exercise is harming them) or if they experience rapid decreases (due to the false belief that they are already cured and do not need to continue)[23]

Adherence and clinical course: A negative correlation was found between pain intensity and exercise frequency in the app (week 8: p=0.004), as well as between frequency of use and improvement in well-being (p=0.01)[21]. Predictive models (LightGBM) confirmed that compliance and motivation outperform demographic variables in predicting pain relief[25].

Psychosocial factors: Clustering analyses suggest that profiles with a high psychiatric burden predict significantly less overall long-term improvement[33]. Likewise, a higher initial symptom burden predicted greater pain interference at one-year follow-up (effect size d=1.27)[26].

In addition, machine learning models have revealed the weight of "protective" variables. It has been discovered that high resilience and the patient's expectation of recovery have a greater predictive potential than classic psychological distress factors (such as depression or anxiety) to identify successful recovery trajectories early[35].

Biomechanical feedback and behavioral engagement: The available evidence suggests that the use of computer vision (Deep Learning) to guide movement not only favors correct technical execution (improving the thickness of the core musculature), but also that real-time audiovisual feedback acts as a key motivational driver. This technology complements professional supervision, making patients feel safe and guided, which is directly associated with higher completion rates of home programs compared with conventional videos[19,36].

Scalability of professional and perceived quality: It has been shown that the use of AI-based clinical support tools enables a drastic increase in attention span (from a ratio of 1:57 to 1:129 patients per physiotherapist) without compromising clinical outcomes related to pain or mental health. On the contrary, the automation of triage and planning optimizes clinicians' time while maintaining identical or even higher levels of adherence and patient satisfaction[24].

4. Discussion

The analysis of the recent literature (2020-2025) reveals a qualitative leap in the application of AI in physiotherapy. While previous reviews characterized the technology as a mere diagnostic observer, current evidence suggests that the systems have integrated assessment, diagnosis, and prognosis, transitioning to an "active therapeutic agent" role. On the one hand, machine learning algorithms have enhanced their prognostic capacity, predicting treatment response, risk of chronicity, and patient adherence with high accuracy[23,26,34,35]. On the other hand, computer vision-based tools enable objective kinematic assessment in real time, facilitating classification into clinical subgroups and the issuance of functional diagnoses[19,30]. It is precisely on this comprehensive clinical basis that current systems can intervene directly through automated exercise prescription and education[19,21,30]. However, a relevant paradox is observed: although the mathematical precision of the models is high, their translation into clinical outcomes (pain and function) is clearly heterogeneous, with trials showing more favorable results of AI-backed interventions compared to conventional treatment[19,20,31,37], others in which only non-inferiority is achieved compared to standard care[28] and studies in which clinical differences barely exceed, or even do not reach the magnitude considered minimally important for the patient[24,31].

A critical inconsistency identified in the literature is the tension between the algorithm's complexity and its clinical utility. It has been detected that "algorithmic opacity" or "Black Box" phenomenon constitutes the main barrier to implementation[10,34]. The inability of deep learning models to explain the reasoning and logic behind a clinical recommendation generates uncertainty and erodes the physiotherapist's confidence[14,24,34]. This suggests that mass adoption will depend less on increasing computing power and more on moving towards "Explainable AI" (XAI), which allows the professional to validate the logic of the decision[11,14,24]. Added to this is technological friction: although adherence is high under experimental conditions, in real practice, operational barriers (technical failures, lack of interoperability) interrupt the clinical workflow[29].

The findings reinforce the need for an "Augmented Intelligence" model versus full automation. Although AI shows potential in data processing and triage (detection of red flags for medical referral and patient stratification of the risk of chronicity to assign the appropriate level of treatment)[3,20], it lacks the affective capacities necessary for therapeutic alliance[14,38]. Therefore, technology is theoretically emerging not as a substitute but as a support that frees the clinician from cognitive and bureaucratic burdens, allowing him to focus on the human aspects of rehabilitation[8,24].

The limitations inherent in this scoping review need to be recognized. The selection of studies, although systematized, is subject to the authors' qualitative interpretation and does not provide a comprehensive quantitative estimate from a meta-analysis. However, this methodological flexibility is justified by the emerging and heterogeneous nature of the technological interventions evaluated, which makes a direct statistical synthesis difficult[1,2]. Likewise, a limitation of external validity in the primary evidence (limited generation) should be considered, since a large part of the algorithms have been trained with data from very specific populations (such as private insurance or high digital literacy), which could perpetuate care biases and make it difficult to generalize the results to public health systems with limited resources or more vulnerable geriatric populations[19,20,25,27].

Another limitation to consider is the very nature of the technology being evaluated. Since Artificial Intelligence is a field in exponential development, it is possible that the models and algorithms analyzed evolve faster than scientific publication cycles. Therefore, there is an inherent temporal obsolescence in this type of revision, since current tools could be quickly surpassed by more advanced systems, which reinforces the need to establish conceptual frameworks for clinical integration that endure beyond specific algorithms.

Although the available evidence places AI as a promising tool for triage and prognosis, its impact on clinical outcomes such as pain and disability remains heterogeneous, and not all interventions consistently achieve the minimum clinically important difference recommended for chronic low back pain[24,28]. This gap suggests that, to improve effectiveness, it is not enough to optimize algorithms; it is necessary to address several structural barriers. At the professional level, training barriers stand out, as many physiotherapists lack digital literacy and the critical interpretation of AI models, which limits their ability to use these tools without delegating their clinical judgment[11,24]. At the systems level, technological barriers persist, such as a lack of interoperability with electronic health records, the instability of some solutions, and the need for specific infrastructure, which add friction to the workflow and reduce adoption outside of experimental settings[3,11]. Finally, ethical and legal barriers stemming from algorithmic opacity and uncertainty about professional responsibility make it difficult for clinicians to trust recommendations generated by "black box" systems[10,34]. Overcoming these barriers would have substantial economic implications for health systems.

From the perspective of cost-effectiveness, some of the included studies provide relevant data: physiotherapy interventions based on computer vision achieved clinical results equivalent to conventional face-to-face therapy, showing to be significantly more cost-effective[30], and chatbot-assisted adherence programs achieved completion rates of 92% with exercise sessions requiring less than one minute per day[37], which suggests a potential for healthcare scalability without a proportional increase in costs. In addition, predictive machine learning models correctly identified, in 83% of the cases evaluated, which patients would not comply with face-to-face treatment before it failed[23], thereby allowing resources to be directed towards more efficient digital interventions for patients with lower adherence. In the Spanish context, where CLBP generates an estimated total cost of almost €9 billion per year[4], even a modest reduction in redundant visits or improved adherence through low-cost digital tools could yield significant healthcare savings. However, these benefits require an initial investment in professional training, interoperable technological infrastructure, and rigorous regulatory evaluation processes, preventing the technology from amplifying existing inequities or transferring the risk of failure to the clinician or patient[10,24].

Based on the identified gaps, specifically the need for clinical safety and scalable personalization[13,24], the conceptual feasibility of "Clinical Copilot" type systems under Human-in-the-loop architecture is proposed. Unlike fully autonomous tools, this theoretical design integrates two sequential modules. First, a strict rule-based filter for Red Flag detection (ensuring medical referral)[13,25,29]. Second, a Natural Language Processing (NLP) engine that enables the system to interpret the patient's free-text descriptions and convert them into structured clinical variables the algorithm can process, eliminating the need for the user to fill out rigid forms. This component is complemented by a retrieval-augmented generation (RAG) architecture. This

technology allows the algorithm to consult a closed database of validated scientific evidence (clinical guidelines) before responding, ensuring that the information is truthful and avoiding false answers or algorithmic "hallucinations". Thus, the system can tailor pain education to the patient's psychosocial profile (e.g., reducing alarmist language in cases of kinesiophobia)[13,14]. The technical architecture of this system (Lumbar-AI), including prompt design and a clinical case simulation, is detailed in Supplementary Texts S1 and S2. This proposal directly addresses the need identified in the review to balance process automation with the indispensable ethical supervision of the physiotherapist[24,38].

5. Conclusions

The review of recent evidence suggests that Artificial Intelligence shows promising results as a clinical support tool, especially in triage and functional prognostic tasks, where machine learning algorithms demonstrate predictive potential comparable to or superior to traditional assessment tools[1,20,35]. However, the direct therapeutic efficacy of these solutions is heterogeneous, positioning the hybrid model (face-to-face intervention powered by digital support) as a pragmatic and promising strategy for managing chronic low back pain[19,22,27,28].

From a conceptual perspective, algorithmic opacity ("Black Box") is one of the most relevant barriers to the clinical integration of these tools, although not necessarily the only or the main one, since it limits the trust necessary for critical decision-making[10,34]. Therefore, the advancement of the field depends largely on a transition towards Explainable AI (XAI) systems that enable professionals to trace the underlying clinical logic[10,24].

As an innovative practical implication, this work suggests that the future of digital care does not lie in independent automation, but in the design of "Human-in-the-loop" "Clinical Co-pilots"[3,24]. These systems have the theoretical feasibility of scaling treatment personalization and red-flag filtering, freeing the physiotherapist from the analytical burden so they can focus on the therapeutic alliance[13,24]. In short, the profession is heading towards a paradigm of "Augmented Intelligence", where the physiotherapist should not maintain a passive user role that leaves control and decision-making to AI, but should lead the ethical and functional design of these new tools to turn them into support instruments that speed up and facilitate clinical practice[2,38]. (See Prompt Architecture and Simulation at Text S1 and S2 of the Supplementary Material)

References

1. Seas A, Zachem TJ, Valan B, Goertz C, Nischal S, Chen SF, et al. Machine learning in the diagnosis, management, and care of patients with low back pain: a scoping review of the literature and future directions. Spine J. 2025; 25(1):18–31. doi: 10.1016/j.spinee.2024.09.010

2. Hasan F, Mudey A, Joshi A. Role of Internet of Things (IoT), Artificial Intelligence and Machine Learning in Musculoskeletal Pain: A Scoping Review. Cureus. 2023; 15(4):e37352. doi: 10.7759/cureus.37352

3. Toh ZA, Berg B, Han QYC, Hey HWD, Pikkarainen M, Grotle M, et al. Clinical Decision Support System Used in Spinal Disorders: Scoping Review. J Med Internet Res. 2024; 26:e53951. doi: 10.2196/53951

4. Alonso-García M, Sarría-Santamera A. The Economic and Social Burden of Low Back Pain in Spain: A National Assessment of the Economic and Social Impact of Low Back Pain in Spain. Spine (Phila Pa 1976). 2020; 45(16):E1026-E1032. doi: 10.1097/BRS.0000000000003476

5. Dal Prete A, Gandolla M, Andreoni G, Braghin F. Low Back Exoskeletons in Industry 5.0: From Machines to Perceiving Co-Pilots—A State-of-the-Art Review. Sensors. 2025; 25(7):1958. doi: 10.3390/s25071958

6. Falla D, Devecchi V, Jiménez-Grande D, Rügamer D, Liew BXW. Machine learning approaches applied in spinal pain research. J Electromyogr Kinesiol. 2021;61:102599. doi: 10.1016/j.jelekin.2021.102599

7. Rajan PV, Eisenberg J, Selverian S, Wawrose R, Yoon ST. The cost-effectiveness of physical therapy versus laminectomy for lumbar spinal stenosis: a Markov decision analysis. Spine J. 2025; 25(8):1608–17. doi: 10.1016/j.spinee.2025.01.005

8. Tropea P, Schlieter H, Sterpi I, Judica E, Gand K, Caprino M, et al. Rehabilitation, the Great Absentee of Virtual Coaching in Medical Care: Scoping Review. J Med Internet Res. 2019; 21(10):e12805. doi: 10.2196/12805

9. Chen K, Zhai X, Sun K, Wang H, Yang C, Li M. A narrative review of machine learning as promising revolution in clinical practice of scoliosis. Ann Transl Med. 2021; 9(1):67. doi: 10.21037/atm-20-5495

10. Ren G, Yu K, Xie Z, Wang P, Zhang W, Huang Y, et al. Current Applications of Machine Learning in Spine: From Clinical View. Global Spine J. 2022; 12(8):1827–40. doi: 10.1177/21925682211035363

11. Jawed AM, Zhang L, Zhang Z, Liu Q, Ahmed W, Wang H. Artificial intelligence and machine learning in spine care: Advancing precision diagnosis, treatment, and rehabilitation. World J Orthop. 2025; 16(8):107064. doi: 10.5312/wjo.v16.i8.107064

12. Kim HB, Kim HS, Kim SJ, Yoo JI. Spine muscle auto segmentation techniques in MRI imaging: a systematic review. BMC Musculoskelet Disord. 2024; 25(1):716. doi: 10.1186/s12891-024-07777-4

13. Luo M, Duan Z, Gao J, Sun Y, Chen L, Feng X. Evaluating the role of ChatGPT in rehabilitation medicine: a narrative review. Front Digit Health. 2025;7:1618510. doi: 10.3389/fdgth.2025.1618510

14. Khan MA, Koh RGL, Rashidiani S, Liu T, Tucci V, Kumbhare D, et al. Cracking the Chronic Pain code: A scoping review of Artificial Intelligence in Chronic Pain research. Artif Intell Med. 2024;151:102849. doi: 10.1016/j.artmed.2024.102849

15. Hill CJ, Banerjee A, Hill J, Stapleton C. Diagnostic clinical prediction rules for categorising low back pain: A systematic review. Musculoskeletal Care. 2023; 21(4):1482–96. doi: 10.1002/msc.1816

16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021; 372:n71. doi: 10.1136/bmj.n71

17. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro Scale for Rating Quality of Randomized Controlled Trials. Phys Ther. 2003; 83(8):713–21. doi: 10.1093/ptj/83.8.713

18. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses [Internet]. Ottawa (ON): Ottawa Hospital Research Institute; [cited 2026 Feb 16]. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

19. Xiao C, Zhao Y, Li G, Zhang Z, Liu S, Fan W, et al. Clinical Efficacy of Multimodal Exercise Telerehabilitation Based on AI for Chronic Nonspecific Low Back Pain: Randomized Controlled Trial. JMIR Mhealth Uhealth. 2025; 13:e56176. doi: 10.2196/56176

20. Cheema C, Baldwin J, Rodeghero J, Werneke MW, Mioduski JE, Jeffries L, et al. Use of machine learning to identify prognostic variables for outcomes in chronic low back pain treatment: a retrospective analysis. J Man Manip Ther. 2025; 33(1):63–72. doi: 10.1080/10669817.2024.2424619

21. Griefahn A, Avermann F, Zalpour C, Marshall RP, Cordon Morillas I, Luedtke K. Exploring the Effect of an 8-Week AI-Composed Exercise Program on Pain Intensity and Well-Being in Patients With Spinal Pain: Retrospective Cohort Analysis. JMIR Form Res. 2025; 9:e57826. doi: 10.2196/57826

22. Alzouhayli K, Schilaty ND, Nagai T, Rigamonti L, McPherson AL, Holmes BD, et al. Artificial intelligence-guided therapy: Clinic versus home users. Long-term differences in patient-reported outcomes in patients with low back pain. J Orthop Rep. 2025; 4(2):100592. doi: 10.1016/j.jorep.2025.100592

23. Pfeifer AC, Schröder-Pfeifer P, Schiltenwolf M, Vogt L, Schneider C, Platen P, et al. Finding Predictive Factors of Stabilization Exercise Adherence in Randomized Controlled Trials on Low Back Pain: An Individual Data Reanalysis Using Machine Learning Techniques. Arch Phys Med Rehabil. 2025; 106(5):738–49. doi: 10.1016/j.apmr.2024.12.015

24. Areias AC, Janela D, Moulder RG, Molinos M, Bento V, Moreira C, et al. Applying AI to Safely and Effectively Scale Care to Address Chronic MSK Conditions. J Clin Med. 2024; 13(15):4366. doi: 10.3390/jcm13154366

25. Areias AC, Moulder RG, Molinos M, Janela D, Bento V, Moreira C, et al. Predicting Pain Response to a Remote Musculoskeletal Care Program for Low Back Pain Management: Development of a Prediction Tool. JMIR Med Inform. 2024; 12:e64806. doi: 10.2196/64806

26. Zhao X, Dannenberg K, Repsilber D, Gerdle B, Molander P, Hesser H. Prognostic subgroups of chronic pain patients using latent variable mixture modeling within a supervised machine learning framework. Sci Rep. 2024; 14(1):12579. doi: 10.1038/s41598-024-62542-w

27. Bates NA, Huffman A, Goodyear E, Nagai T, Rigamonti L, Breuer L, et al. Physical clinical care and artificial-intelligence-guided core resistance training improve endurance and patient-reported outcomes in subjects with lower back pain. Clin Biomech. 2023;103:105902. doi: 10.1016/j.clinbiomech.2023.105902

28. Marcuzzi A, Nordstoga AL, Bach K, Aasdahl L, Nilsen TIL, Bardal EM, et al. Effect of an Artificial Intelligence–Based Self-Management App on Musculoskeletal Health in Patients With Neck and/or Low Back Pain Referred to Specialist Care: A Randomized Clinical Trial. JAMA Netw Open. 2023; 6(6):e2320400. doi: 10.1001/jamanetworkopen.2023.20400

29. Tun Firzara AM, Teo CH, Teh S, Su JY, Zaini HSM, Suhaimi AB, 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):742–52. doi: 10.1093/fampra/cmad044

30. Park C, Yi C, Choi WJ, Lim HS, Yoon HU, You SJH. Long-term effects of deep-learning digital therapeutics on pain, movement control, and preliminary cost-effectiveness in low back pain: a randomized controlled trial. Digit Health. 2023;9:20552076231217816. doi: 10.1177/20552076231217816

31. Hartmann R, Avermann F, Zalpour C, Griefahn A. Impact of an AI app-based exercise program for people with low back pain compared to standard care: A longitudinal cohort-study. Health Sci Rep. 2023; 6(1):e1060. doi: 10.1002/hsr2.1060

32. Nordstoga AL, Aasdahl L, Sandal LF, Dalager T, Kongsvold A, Mork PJ, et al. The Role of Pain Duration and Pain Intensity on the Effectiveness of App-Delivered Self-Management for Low Back Pain (selfBACK): Secondary Analysis of a Randomized Controlled Trial. JMIR Mhealth Uhealth. 2023; 11:e40422. doi: 10.2196/40422

33. Strigo IA, Simmons AN, Giebler J, Schilling JM, Moeller-Bertram T. Unsupervised learning for prognostic validity in patients with chronic pain in transdisciplinary pain care. Sci Rep. 2023; 13(1):7603. doi: 10.1038/s41598-023-34611-z

34. Zmudzki F, Smeets RJEM. Machine learning clinical decision support for interdisciplinary multimodal chronic musculoskeletal pain treatment. Front Pain Res. 2023;4:1177070. doi: 10.3389/fpain.2023.1177070

35. 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(1):834. doi: 10.1186/s12891-022-05718-7

36. Park J, Chung S, Park JH. Real-Time Exercise Feedback through a Convolutional Neural Network: A Machine Learning-Based Motion-Detecting Mobile Exercise Coaching Application. Yonsei Med J. 2022; 63 Suppl:S34–42. doi: 10.3349/ymj.2022.63.S34

37. Anan T, Kajiki S, Oka H, Fujii T, Kawamata K, Mori K, et al. Effects of an artificial intelligence-assisted health program on workers with neck/shoulder pain/stiffness and low back pain: Randomized controlled trial. JMIR Mhealth Uhealth. 2021; 9(9):e27535. doi: 10.2196/27535

38. Soer R. Artificial intelligence in musculoskeletal rehabilitation. J Back Musculoskelet Rehabil. 2023; 36(5):993–4. doi: 10.3233/BMR-235003

Editorial information

Correspondence

869739@unizar.es

Dates

Received: 08.03.2026
Accepted: 02.06.2026
Published: 23.07.2026

Author Contributions

Conceptualization, M.C.L.; methodology, M.C.L.; research, M.C.L.; writing—preparation of the original draft, M.C.L.; writing—revision and editing, M.C.L.; supervision, M.C.L. All authors have read and accepted the published version of the manuscript.

Funding

This research did not receive external funding.

Data Availability Statement

The data presented in this study are available in the body of the article and in the supplementary material. No primary data were generated with patients, since the research consists of a bibliographic review and a theoretical proposal

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Suggested citation

Campillos Lázaro M. Use of Artificial Intelligence as a Clinical Decision Support System in the Management of Chronic Low Back Pain. A Scoping Review with Applied Conceptual Component. Invasive Physiother Musculoskelet Med. 2026;2:e25. doi: 10.63360/ipmm.v2.e25

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of IPMM and/or the editor(s). IPMM and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Copyright

© 2026 by the authors.

Publication under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA) license (https://creativecommons.org/licenses/by-nc-sa/4.0)

Supplementary Material
Query Date and Filters

Date: November 21-22, 2025.

Filters: Publications 2020-2025; English or Spanish language.

1. Search Strategy

To avoid repetitions, three thematic blocks were defined combined using Boolean operators:

Block A (Technology): "Artificial Intelligence" OR "Decision Support Systems" OR "Machine Learning" OR "Deep Learning" OR "Neural Networks" OR "ChatGPT" OR "Large Language Model" OR "Generative AI" OR "CDSS" OR "Chatbot" OR "Virtual Assistant". (En PubMed se añadieron los términos MeSH correspondientes).

Block B (Intervention): "Physical Therapy" OR "Physiotherapy" OR "Rehabilitation" OR "Exercise Therapy".

Block C (Patología): "Low Back Pain" OR "Chronic Pain" OR "Sciatica" OR "LBP" OR "Lumbar" OR "Back Pain".

Table S1. Detailed search strategy and results by database.

Database	Syntax and Specific Fields	Results
PubMed (MEDLINE)	(Bloque A [MeSH/All Fields]) AND (Bloque B [MeSH/All Fields]) AND (Bloque C [MeSH/All Fields])	146
Scopus	TITLE-ABS-KEY(Block A) AND TITLE-ABS-KEY(Bloque B) AND TITLE-ABS-KEY(Block C)	264
Web of Science (WoS)	TS=(Bloque A AND Block B AND Bloque C). Field: Topic.	170
Cochrane Library	Advanced Search: Line #1 (Block A) AND Line #2 (Block B) AND Line #3 (Block C)	55
PEDro	Simple search by keyword term "Artificial Intelligence" (with subsequent manual screening).	26
GROSS TOTAL	Sum of identified records (before removing duplicates)	661

2. Evaluation of Methodological Quality

Table S2. PEDro scale score breakdown for included clinical trials.

Author and Year	Total	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Park et al., 2023	7/10	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓
Xiao et al., 2025	7/10	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓
Tun Firzara et al., 2023	7/10	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓
Marcuzzi et al., 2023	6/10	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓
Nordstoga et al., 2023	6/10	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓
Anan et al., 2021	4/10	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓
Bates et al., 2023	3/10	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓

PEDro Criteria Legend:

C1: Specified Choice Criteria (does not add to the total).

C2: Random assignment.

C3: Hidden Assignment (Masking).

C4: Initial similarity between groups.

C5: Blinding of subjects.

C6: Blinding of therapists.

C7: Blinding of evaluators.

C8: Adequate follow-up (>85%).

C9: Intention-to-treat analysis.

C10: Statistical comparison between groups.

C11: Specific and variability measures.

Methodological Quality Assessment (PEDro Scale)

The methodological quality of the 7 included randomised clinical trials (RCTs) had a mean score of 5.7/10 (range: 3-7). The vast majority of studies (n=5) demonstrated high methodological quality (≥6/10), standing out for strict compliance with random assignment (C2) and rigorous statistical reporting (C10, C11), as well as high adherence to intention-to-treat analysis (C9, present in 4 of these 5 trials). In contrast, two trials (Anan et al. and Bates et al.)[27,37] were at high risk of bias (≤4/10) due to significant deficiencies in allocation concealment and high attrition rates.

The cross-sectional analysis of the items reveals that the main methodological weakness of the current literature is blinding. As a difficulty inherent in active exercise and digital technology interventions, the blinding of the therapists (C6) was nil and that of the subjects (C5) was almost anecdotal (present in a single study). However, this risk of bias was partially mitigated by the fact that blinding by independent assessors (C7) was successfully implemented in more than half of the trials (n=4).

Table S3. Newcastle-Ottawa Scale (NOS) score breakdown for included cohort and case-control studies.

Author and Year	Total	Design	S1	S2	S3	S4	C1	C2	R1	R2	R3
Areias et al. (Tool), 2024	9/9	Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pfeifer et al., 2025	9/9	Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓
Zmudzki & Smeets, 2023	8/9	Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✗
Zhao et al., 2024	8/9	Cohort	✓	✓	✓	✗	✓	✓	✓	✓	✓
Areias et al. (Scale), 2024	8/9	Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✗
Cheema et al., 2024	7/9	Cohort	✗	✓	✓	✓	✓	✓	✓	✓	✗
Knoop et al., 2022	7/9	Cohort	✓	✓	✗	✓	✓	✓	✗	✓	✓

Author and Year	Total	Design	S1	S2	S3	S4	C1	C2	R1	R2	R3
Strigo et al., 2023	6/9	Cohort	✓	✓	✓	✓	✓	✗	✗	✓	✗
Alzouhayli et al., 2025	4/9	Cohort	✓	✗	✓	✓	✗	✗	✗	✓	✗
Griefahn et al., 2025	4/9	Cohort	✓	✗	✓	✓	✗	✗	✗	✓	✗
Hartmann et al., 2023	4/9	Cohort	✗	✗	✓	✓	✗	✗	✗	✓	✓
Park et al., 2022	3/9	Cases and controls	✗	✓	✓	✓	✗	✗	✗	✗	✗

Newcastle-Ottawa Criteria Legend (NOS): Each item is scored with a maximum of 1 point (✓), except for comparability which allows up to 2 points. Maximum score: 9.

Selection (S1-S4):

- S1: Representativeness of the exposed cohort.
- S2: Selection of the unexposed cohort.
- S3: Determination of exposure.
- S4: Demonstration that the result was not present at the beginning.

Comparability (C1-C2):

- C1: Control of the main confounding factor (e.g., age/severity).
- C2: Control of additional factors.

Results (R1-R3):

- R1: Independent/blind outcome evaluation.
- A2: Long enough tracking.
- R3: Adequacy of follow-up (losses <20%).

Methodological Quality Assessment (Newcastle-Ottawa Scale)

The evaluation of the 12 observational studies (11 cohorts and 1 case-control) using the Newcastle-Ottawa Scale (NOS) revealed heterogeneous methodological quality (range: 3-9 out of 9 possible points). Almost half of the sample (n=5) demonstrated high quality (≥8/9), highlighting the works of Areias et al. (Tool)[25] and Pfeifer et al.[23] with perfect scores, which exhibited great robustness in the representativeness of the cohort, multivariate statistical adjustment and evaluation of results. At an intermediate level, three studies showed a moderate risk of bias (6-7/9), with minor deficiencies in demonstrating no result at baseline or in blinded assessment.

In contrast, one-third of the investigations (n=4) were at high risk of bias (≤4/9). Cross-sectional analysis of the domains shows that comparability was the main limiting factor: the lower quality studies consistently failed to control for key confounding factors (such as age or baseline severity). Likewise, within the Results domain, the adequacy of monitoring was identified as the main overall weakness; More than half of the studies (n=7) reported significant losses or did not adequately detail retention rates, introducing potential attrition bias into the accumulated evidence.

3. Technical Architecture of the System

Text S1. Lumbar System Prompts Architecture-AI.

Prompt Level 1: Security Module

ROLE: Physiotherapist Expert in triage (Primary Care).

TASK: Analyze Red Flags in the patient's input.

CRITICAL CONSTRAINTS:

If there is saddle anesthesia, loss of bowel control, fever/unexplained weight loss, or severe trauma -> ACTION: "STOP_URGENCIA."

If constant, non-mechanical and nocturnal pain -> ACTION: "STOP_DERIVACION_MEDICA".

Only if there are NO red flags -> ACTION: "PASS" and classify phenotype.

Phenotypes:

1: Simple mechanical nociceptive (low psychosocial risk)

Clinical profile: Low back pain associated with movement or posture (e.g., flexion/extension intolerance or stability deficiency), no red flags. The patient presents high self-efficacy, resilience and absence of significant emotional distress. Impact on Intervention Module: Activates a direct motor control and strengthening protocol, with emphasis on mobility and progressive exercise. The educational capsule focuses on basic self-management and promoting adherence to exercise, given the expected good prognosis

2: Mechanical nociceptive with kinesiophobia (fear-avoidance)

Clinical profile: Mechanical pain associated with maladaptive beliefs of tissue damage. High scores in fear of movement (e.g., TSK-11) and physical activity avoidance behaviors. This phenotype corresponds to the simulation case described in Text S2. Impact on Intervention Module: Activates a gradual exposure protocol, selecting exercises with a low initial biomechanical load to reintroduce movement. The educational capsule prioritizes the restructuring of the belief "pain = harm" and the use of non-threatening language

3: Mechanical Nociceptive with Emotional Distress

Clinical profile: Pain of mild-moderate intensity that is accompanied by high perceived interference due to high levels of anxiety, low mood, or poor social support. It corresponds to patients who, despite not presenting the greatest physical pain, show a high functional impact due to emotional factors. Impact on Intervention Module: The algorithm prioritizes the inclusion of stress regulation components (breathing, relaxation, mindfulness) combined with exercise. The educational capsule focuses on stress management, normalizing chronic pain, and improving coping

4: High irritability / global severity (mixed/nociplastic pattern)

Clinical profile: High-intensity pain, relevant disability (e.g., elevated ODI), possible extension to multiple areas, and high levels of catastrophizing and distress. Although the patient is "FIT"

(without red flags), he or she has signs of central sensitization and a high risk of poor outcome. Impact on Intervention Module: Activates a conservative protocol with exercises focused on gentle mobility in painless ranges and intensive pain education. The system may include an internal indicator to suggest multidisciplinary assessment (physiotherapy + psychological approach) if the clinical response is insufficient after a reasonable period of time

EXPECTED OUTPUT: JSON {"Status": "PASS/STOP", "Razon_Seguridad": "...", "Phenotype": "1/2/3/4" (if PASS), "Driver_Principal": "Mechanical / Kinesiophobia / Stress / Severity"}.

Prompt Level 2: Intervention Module

CONTEXT: Patient with Chronic Low Back Pain classified "PASS".

OBJECTIVE: To design a home exercise session (20 min).

CUSTOMIZATION INSTRUCTIONS:

Selection: Based on motor control and gradual exposure.

Education: ALWAYS include "Educational Capsule" (pain ≠ harm).

Psychosocial: If high kinesiophobia (>40 TSK-11), use non-threatening language (avoid words such as "damage", "breakage", "wear and tear").

EXPECTED OUTPUT: Table with 3 exercises + 1 Educational Council.

4. Proof of Concept

Text S2. Simulation of Clinical Case.

To illustrate how the *Lumbar-AI Assist* system works in a controlled environment, a complete simulation with a standard patient profile is presented.

1. INPUT PROFILE

Patient: Male, 45 years old. Administrative.

Reason for consultation: Low back pain of 3 months of evolution, exacerbated when sitting.

Scales:

EVA: 6/10.

Oswestry (ODI): 24% (Moderate disability).

TSK-11 (Kinesiophobia): 48/68 (High).

Red flags: negative.

2. ALGORITHM PROCESSING

Module 1 (Security): Detects the absence of alarm symptoms. Classifies as PASS.

Presumptive phenotype: Mechanical nociceptive with psychosocial component (fear-avoidance) (Phenotype 2 / Driver: Kinesiophobia)

Module 2 (Intervention): Activates the "Gradual Exposure" protocol due to high kinesiophobia (TSK > 40). Select exercises with a low initial biomechanical load.

3. SYSTEM OUTPUT (OUTPUT GENERATED)

Educational Capsule:

Message generated: "John, your back is sturdy and secure. The pain when sitting is a 'sensitive alarm', not real harm. Gentle movement is the best pain reliever."

Prescription of Tasks:

Supine Pelvic Scale (3x10): To reintroduce movement without fear.

Bipodal Gluteal Bridge (3x8, comfort range only): For low-load strength.

Cat-Camel (2 min continuous, slow): For motor control.

Progression Criteria: If post-exercise pain is < 4/10, suggest increasing the load in the next session.