



OPEN

A data-driven analysis of lumbar steroid injection satisfaction in patients with chronic low back pain

Maria Monzon^{1,2}✉, Iara De Schoenmacker¹, Andrea Cina^{1,3}, Réka Enz^{1,4}, Christian Lanz⁴, Fabio Galbusera³, Catherine R. Jutzeler^{1,2,5} & Zina-Mary Manjaly^{1,4,5}

Chronic low back pain (CLBP) is a prevalent condition significantly reducing quality of life. Lumbar steroid injections are a widely used conservative treatment option, but their effectiveness varies among patients. This study aimed to develop a predictive framework that integrates clinical variables and patient demographics to evaluate post-treatment pain satisfaction in CLBP patients undergoing lumbar injection therapy. We performed a retrospective analysis of 212 CLBP patients to evaluate the treatment satisfaction and pain intensity changes using the Numerical Rating Scale (NRS). A Random Forest model, validated through nested cross-validation, achieved an average precision of 0.865 in predicting treatment satisfaction. SHapley Additive exPlanations (SHAP) analysis revealed pain self-efficacy features, particularly coping mechanisms and household activities, as key outcome predictors of post-treatment pain satisfaction. Clinically significant pain reduction thresholds were identified at an absolute change of 2.09 and a relative change of 30 % on the NRS. Our findings reveal the biological and social factors influencing post-treatment pain in CLBP patients. The identified pain reduction thresholds and predictors may help clinicians to develop individualized management strategies, optimizing treatment outcomes and improving patient care. Future research should refine the predictive model by incorporating additional multimodal variables to better capture CLBP heterogeneity.

Low back pain (LBP) is a highly prevalent condition that affects more than 600 million people worldwide¹, with a lifetime prevalence of up to 80%^{2–4}. Approximately 10% of the cases persist for over 3 months, meeting the criteria for chronic low back pain (CLBP). The rising prevalence of CLBP⁵ poses significant challenges for individuals and public health systems⁶, as it remains the leading cause of years lived with disability worldwide^{1,7}. Low back pain, defined as discomfort between the costal margins and the inferior gluteal folds, is often accompanied by leg pain and may present with additional symptoms such as stiffness, reduced range of motion, muscle spasms, localized tenderness, paresis, numbness, or tingling⁸. It can result from various causes, including nerve injury, spinal cord compression, muscle or ligament damage, inflammation, or infection^{8,9}. The location and characteristics of pain can provide clues to its etiology, but identifying the precise source and determining the optimal treatment in clinical practice often requires considerable trial and error. This process is further complicated by the fact that the etiology of CLBP is frequently linked to psychosocial factors, with patients commonly reporting symptoms such as poor concentration, disrupted sleep, memory difficulties, and irritability¹⁰. The biopsychosocial model of pain¹¹ describes the complex interplay among biological, psychological, and social factors contributing to the pathophysiological heterogeneity of CLBP. This complexity likely underlies and explains the considerable variability in treatment effectiveness¹².

Lumbar injection, which involve administering local anesthetics and steroids to structures of the lumbar spine¹³, is a therapeutic option for patients with CLBP who do not respond to first-line analgesics and physical therapy. Despite its widespread use, evidence regarding the effectiveness of infiltration therapy for CLBP remains inconsistent, with studies reporting mixed outcomes across patients^{14,15}. While some patients experience significant symptom relief, others gain little or no benefit, possibly due to differences in pain perception, psychological factors, and comorbidities^{16,17}. Recognizing this variability, previous research emphasizes the need to understand the mechanisms driving treatment outcomes and develop individualized treatment strategies^{17–19}.

¹Department of Health Sciences and Technology (DHEST), ETH Zurich, 8092 Zürich, Switzerland. ²Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland. ³Department of Teaching, Research and Development, Schulthess Clinic, 8008 Zürich, Switzerland. ⁴Department of Neurology, Schulthess Clinic, 8008 Zürich, Switzerland. ⁵These authors jointly supervised this work: Catherine R. Jutzeler and Zina-Mary Manjaly. ✉email: maria.monzonronda@hest.ethz.ch

Personalized care approaches, using prognostic profiling and clinical prediction models, have demonstrated potential to improve treatment outcomes^{20,21}. Although predictive tools show potential in identifying patients who could benefit from infiltration therapy²², their clinical implementation and utility remain very limited²³. A significant barrier to progress is the absence of a well-defined outcome measure tailored specifically to assess the success of lumbar injection therapy²⁴.

Currently, the severity of CLBP and its associated disability are commonly assessed using patient-reported outcome measures (PROMs)^{22,25,26}, such as the numerical rating scale (NRS)²⁷ and visual analog scale (VAS)²⁷ for pain intensity. Establishing a threshold that represents the smallest change in PROM scores perceived as beneficial by patients is essential to determine the clinical significance of a therapy^{28,29}. While previous studies have explored such thresholds in surgical contexts^{30–33}, there is a lack of studies addressing this need for lumbar steroid injection therapy. To bridge this gap, we leveraged data-driven methodologies to develop a comprehensive predictive framework for lumbar injection therapy in patients with CLBP by integrating clinical data and patient-specific demographics. First, we developed a predictive model to identify key factors influencing the effectiveness of lumbar steroid injection therapy, enabling the identification of patients most likely to experience improvements in pain perception. Treatment success was evaluated based on self-reported pain satisfaction following therapy. Next, we aimed to establish clinically relevant thresholds for pain reduction specific to infiltration therapy, focusing on the minimal reduction in pain scores required for patients to perceive treatment outcomes as satisfactory.

Methods

Study design and participants

A retrospective secondary analysis was performed using data from the *Treatment Expectation and their Influence on Infiltration outcome* (TREXI) study¹⁶. The TREXI study was a prospective observational longitudinal investigation carried out between February 2019 and December 2020 at the Department of Neurology, Schulthess Clinic in Zurich, Switzerland. The original cohort included 306 adult patients, aged 18 to 93 years, diagnosed with CLBP. For this secondary analysis, a subset of 212 patients who provided informed consent for the additional use of their data in research was included (Fig. 1a). The study was approved by the Cantonal Ethics Committee of Zurich (BASEC-NR 2023-02210) and complied with the ethical principles outlined in the Declaration of Helsinki.

Our study focused on patient-specific clinical and demographic characteristics as potential predictors of treatment response, excluding measures of patient expectations that were the main focus of the original analysis¹⁶. The experimental protocol comprised questionnaires administered in German at three time points: on the day receiving the lumbar steroid injection, immediately prior to treatment, immediately after receiving the lumbar steroid injection and two weeks after the treatment.

To align with the study's objective of evaluating predictors of treatment response, data collected immediately after the treatment were excluded. This exclusion was implemented to ensure a clear separation between the baseline data and the post-treatment data. The baseline was redefined as T_0 , representing the period prior to injection therapy, and the post-treatment period was labeled as T_1 , corresponding to two weeks after injection.

Measures

Data collection encompassed a comprehensive set of questionnaire items addressing patients' demographics, pain characteristics, and self-reported health status (Figure 1b).

Demographics: Demographic information including age, sex, and education level, as well as categories of professional status — categorized as self-employed, student, homemaker, retired, incapacitated, or unemployed — was collected through questionnaires.

Pain characteristics: The duration of back complaints was recorded into intervals, namely less than 4 weeks, 4 to 8 weeks, 8 to 12 weeks, and more than 12 weeks. Current back pain was assessed using numerical rating scales (NRS)²⁷ which involve individuals rating their pain intensity on a scale from 0 (no pain) to 10 (worst pain imaginable). For participants who had previously undergone lumbar steroid injections, additional data were collected, including whether they experienced improvement after the last injection, the time elapsed since the previous treatment — categorized as less than 1 year, 1 to 2 years, or more than 2 years — and whether the procedure was performed by the same doctor or clinic. Motivation for treatment was assessed through sources of influence, including friends, family, the doctor performing the infiltration, general practitioner, internet, personal experience, and the importance of others' opinions.

Self-reported health status: A comprehensive set of validated PROMs covering medication beliefs, expectations, empathy in care, and self-efficacy were collected when infiltration therapy was administered (T_0) to gain a multidimensional understanding of the patient's pain experience and its impact. The items were extracted from widely used questionnaires in clinical research²⁶ and consisted of the following: The Perceived Sensitivity to Medicine (PSM) scale was utilized to evaluate patients' perceived responsiveness to medication in general³⁴. This questionnaire includes items assessing perceived susceptibility to medications, beliefs about experiencing strong reactions, perceptions of having stronger reactions than others, and concerns about side effects from regular medication use. Responses were recorded on a 5-point Likert scale, ranging from "strongly disagree" to "strongly agree." Furthermore, the Consultation and Relational Empathy (CARE) measure was employed to record their evaluation of the overall care experience³⁵. This questionnaire assesses various aspects of the patient-provider interaction, such as the provider's ability to make the patient feel at ease, allow them to tell their story, feeling understood by the healthcare provider, be interested in them as a whole person, fully understand their concerns, show care and compassion, and explain things clearly. Responses are given on a 5-point Likert scale ranging from "poor" to "excellent". For multidimensional assessment of pain and disability, the Core Outcome Measures Index (COMI) back score was used^{36,37}. The COMI questionnaire is a concise 7-item questionnaire which

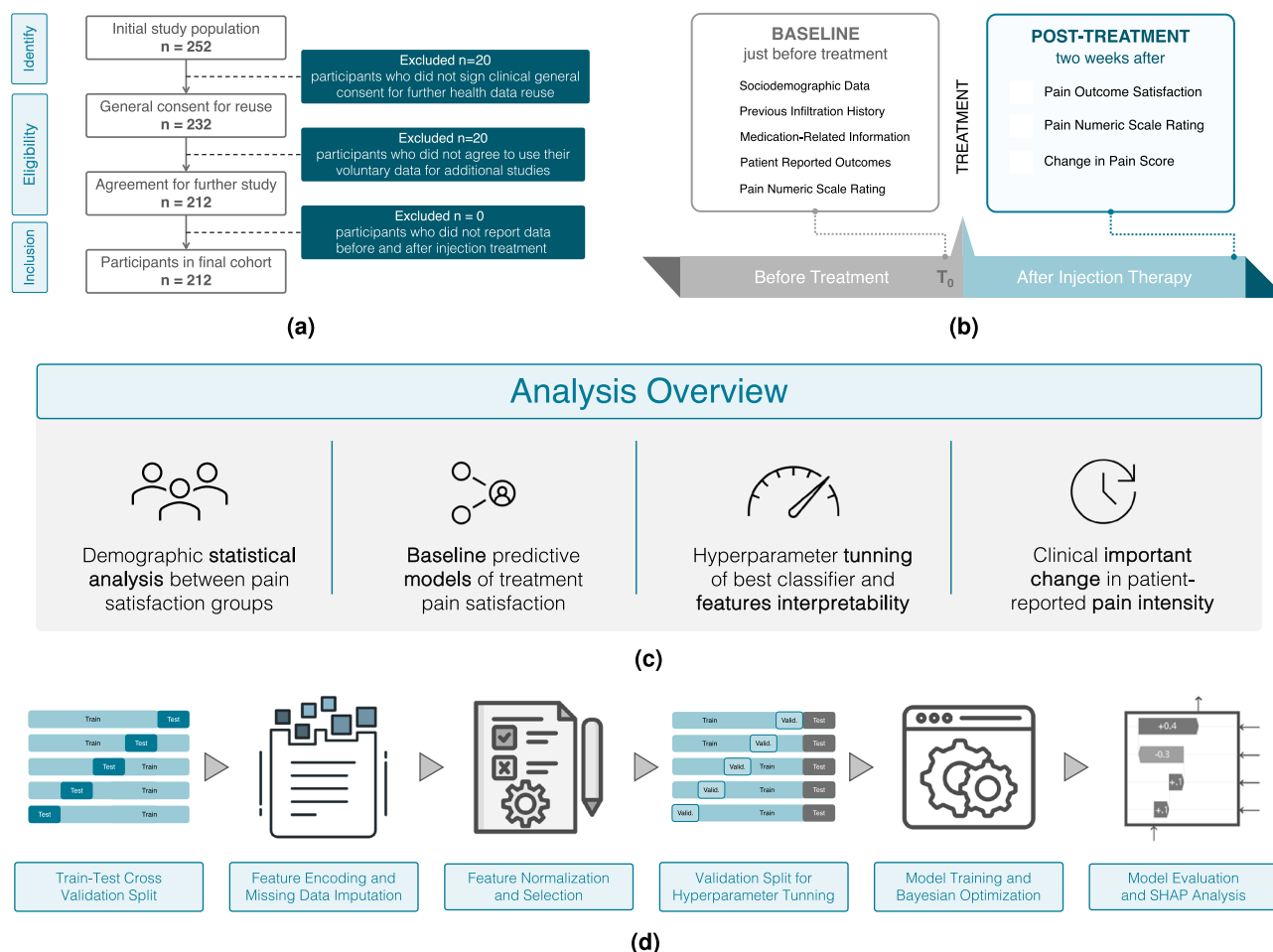


Fig. 1. Study design for predicting patient satisfaction with lumbar steroid injection therapy. **(a)** Retrospective cohort selection: the flowchart illustrates the inclusion procedure of participants for the retrospective cohort analysis. **(b)** Data collection timeline for evaluating lumbar steroid injection outcomes in chronic low back pain patients, including baseline assessments (T_0) and two-week post-treatment follow-up (T_1). **(c)** Statistical analysis framework comprising demographic variable analysis and machine learning baseline predictive modeling, classifier optimization incorporating feature selection, model hyperparameter tuning and feature importance (SHAP) analysis, and ROC curve analyses for clinical significance of minimal change in reported pain metrics. **(d)** Predictive model development: the process starts with nested cross validation, splitting data for model evaluation and hyperparameter tuning. Data preprocessing includes managing missing values, encoding categorical variables, and scaling features. Feature engineering includes feature normalization and selection, where the most informative ones are chosen based on statistical tests. Model tuning in the inner loop optimizes the best baseline classifier hyperparameters. The model's performance is assessed with metrics such as AUC, F1-score, and Average Precision. Finally, SHAP analysis interprets predictions and identifies key features influencing patient satisfaction.

assesses the LBP disability, quality of life and pain perception including questions on back and leg pain intensity (0–10 NRS scale), function, symptom-specific well-being, general quality of life, and disability at work and social situations (5-point Likert scale)^{36,38}. The COMI has been extensively validated^{39–42} against well established longer questionnaires such as the Roland Morris Disability Questionnaire^{37,43} and 36-item short-form health survey (SF-36)^{44,45}. The Pain Self-Efficacy Questionnaire (PSEQ) was used to evaluate patients' beliefs in their ability to cope with and manage pain despite its presence⁴⁶. This questionnaire includes 10 items assessing the patient's confidence in performing various activities despite pain, such as enjoying things, doing household chores, socializing, coping with pain without medication, achieving goals, engaging in leisure activities, coping with pain in general, accomplishing work tasks, leaving a normal lifestyle, and becoming more active. Responses are provided on a 7-point scale ranging from 0 ("not at all confident") to 6 ("completely confident").

Outcomes

The primary focus of this study was to assess the effectiveness of lumbar steroid injections by evaluating both patient satisfaction and clinically meaningful improvements in pain intensity levels two weeks after treatment (T_1). Accordingly, the main outcome was *pain level satisfaction*. A dichotomized variable was created based

on the question: “Are you satisfied with the current pain level?” (“Sind Sie mit dem aktuellen Schmerzniveau zufrieden?”). Patients rated their satisfaction on a scale from 0 to 10, where 0 indicated complete satisfaction and 10 indicated no satisfaction at all. To align with clinical success criteria, a cut-off point of 6 was established. Patients scoring between 0 and 6 were classified as satisfied, reflecting a successful treatment outcome, while those scoring between 7 and 10 were classified as dissatisfied. Secondary outcomes focused on changes in pain intensity to objectively assess improvements in maximum pain levels. The baseline pain (NRS_{T_0}) level was computed as the maximum pain reported in the first two questions of the COMI questionnaire referring to back and leg pain evaluated using 0–10 NRS. The absolute change in pain (ΔPain) was calculated as the difference between baseline pain (NRS_{T_0}) and pain two weeks after treatment (NRS_{T_1}), computed as the maximum between back and leg pain reported at T_1 . To account for individual variability in baseline pain levels, a relative change in pain ($\Delta_r\text{Pain}$) was calculated by normalizing the absolute change to the baseline value, as follows:

$$\Delta_r\text{Pain} = \frac{NRS_{T_0} - NRS_{T_1}}{NRS_{T_0}}$$

Statistical analysis

Figure 1c illustrates the comprehensive statistical analysis workflow, from initial data preprocessing through descriptive statistics to feature analysis of predictive models. Descriptive statistics were used to summarize the general characteristics of the participants across the two groups created according to the “pain level satisfaction” variable.

Continuous variables were reported as mean \pm standard deviation. Categorical variables were presented as frequencies and percentages. For groups comparison, Student’s t-test (or Wilcoxon signed-rank tests when appropriate) and chi-square test were used for continuous and categorical variables, respectively. To control for multiple comparisons and maintain a false discovery rate of 5%, all statistical comparisons were adjusted using the Benjamini-Hochberg correction method. All statistical analysis were performed using the statsmodels library in Python version 3.10.

Baseline predictive models of treatment satisfaction

A data-driven predictive model was developed to classify treatment outcomes based on the dichotomized pain level satisfaction variable, distinguishing between “satisfied” and “dissatisfied” patients. This section outlines the methodological approach used to design and benchmark predictive models. The development and implementation of the predictive models, were performed using Python version 3.10, with the scikit-learn and PyCaret⁴⁷.

Outer cross-validation data split

To mitigate the potential risk of overfitting associated with the limited sample size, a stratified nested cross-validation (CV) methodology was employed⁴⁸. Nested CV involves two iteration loops over the data. In the first iteration, the outer loop applied a 10-fold stratified CV scheme to divide the dataset into training and testing sets, ensuring class balance (i.e., satisfied and dissatisfied) across folds and providing unbiased estimates of model generalization performance on completely unseen data.

Feature engineering and data preprocessing

Prior to predictive model training, several preprocessing steps were performed to ensure the quality and relevance of the features. Features with missing values exceeding 15% were removed, given the limited dataset size, to prevent bias and ensure reliable analysis. The remaining missing data were imputed using an iterative approach: Random Forest (RF) was used for numeric features, and K-Nearest Neighbors (KNN) for categorical features, iterating five times for optimal imputation. Importantly, data imputation was completed prior to the outer cross-validation split to avoid any information leakage. Features were encoded based on their data type to ensure effective model training: numeric features were standardized using z-scores for consistent scaling, categorical variables were one-hot encoded to convert them into a numerical format, and ordinal features were label encoded to preserve their inherent order. To address potential multicollinearity and improve classification performance, features with a variance less than 0.01, as well as those with a correlation coefficient greater than 0.7, were removed.

Baseline models training

During each outer loop cycle, the training data was further split using a 10-fold stratified inner CV to train the classifier. This step ensured the selection of optimal model configurations without introducing information leakage from the test set. This nested approach kept model training and hyperparameter tuning separate from the final performance evaluation, thereby improving the reliability of the generalization assessments.⁴⁹

Multiple baseline classifiers were trained on all the features and benchmarked using various classification algorithms: Logistic Regression⁵⁰ (LR), KNN⁵¹, Support Vector Machines (SVM)⁵² with linear kernel⁵², Ridge Classifier⁵³ (RC), Naive Bayes⁵⁴ (NB), Linear (LDA) and Quadratic Discriminant Analysis⁵⁵ (QDA), Decision Trees⁵⁶ (DT), and ensemble methods including Extra Trees⁵⁷ (ET), RF⁵⁸, AdaBoost⁵⁹, Gradient Boosting Machine⁶⁰ (GBM), XGBoost⁶¹, and LightGBM⁶². These baseline comparison provided a reference point for subsequent optimization.

Classification model performance evaluation

Predictions from all outer loop iterations were concatenated to calculate the final performance metrics, providing a robust estimate of the model’s generalization ability while maintaining strict train-test separation.

Model performance was primarily assessed based on F1-score, average precision (AP), and Matthews correlation coefficient (MCC), which are classification metrics particularly useful for imbalanced datasets. The F1-score represents the harmonic mean of precision and recall calculated as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + 0.5 \times (FP + FN)}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F1 score ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0.5 indicating random guessing. The AP for a given class is calculated as the area under the precision recall curve:

$$AP = \sum_{n=1}^N P(n) \Delta R(n)$$

where $P(n)$ denotes precision at the n -th recall level, while $\Delta R(n)$ measures changes between consecutive recall levels. AP ranges from 0 to 1, with 1 indicating optimal precision and recall and 0 indicating performance equivalent to random guessing. MCC is a robust summary metric computed as the correlation coefficient between observed and predicted binary classifications⁶³:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC values range from -1 to +1, where +1 represents a perfect prediction, 0 indicates that the prediction is not better than a random prediction, and -1 represents total disagreement between prediction and observation. In addition, standard classification metrics including precision, recall, accuracy, Cohen's Kappa coefficients, and Area under the Receiver Operating curve (AUC) were used for a comprehensive evaluation.

Optimized predictive models of treatment satisfaction

Building on the baseline comparison, the best-performing model from the cross-validation splits was further tuned (Fig. 1d) with the objective of identifying the most informative clinical and demographic characteristics through statistical techniques. The data encoding procedures, i.e., encoding and standardization of variables, remained consistent with the approach used in the baseline classifiers, ensuring methodological uniformity.

Feature selection and oversampling

Next, feature selection was performed according to an XGBoost estimator based importance ranking, retaining the top 70% (35 features). To address the class imbalance in the dataset, random oversampling was applied during model training. This technique involved duplicating samples from the minority class to balance the class distribution.

Optimized model training and hyperparameter optimization

We optimized hyperparameters for the best-performing baseline model, RF, using Bayesian and random grid search. This optimization aimed to identify the most informative clinical and demographic characteristics through systematic exploration of the model's parameter space while minimizing overfitting. The hyperparameters were tuned through 5-fold stratified cross-validation (inner folds) for each outer fold, with predefined hyperparameter space ranges.

The RF model was configured with 10 to 1000 trees, with a higher number of trees that potentially improve performance, but increase computational time. The tree depth ranged from 1 to 32, allowing the model to capture more complex patterns, although deeper trees carry a higher risk of overfitting. The risk of overfitting was mitigated by adjusting split node samples (2 to 20) and leaf samples (1 to 20). The feature fractions for splitting were varied between 0.1 and 1.0, with smaller values introducing randomness to reduce overfitting.

Interpretability

Feature importance was assessed using SHapley Additive exPlanations (SHAP), which identified the principal predictors for the best-performing classifier. SHAP values provide a quantitative measure of the influence of individual features, representing the average marginal contribution of each feature to the model's prediction for a given instance⁶⁴. These values are computed by comparing the model's predictions with and without each feature, considering all possible feature combinations⁶⁴. Larger absolute SHAP values indicate stronger effects, while the sign of the value shows whether a feature with a positive SHAP value increases or decreases the prediction outcome.

Clinical important absolute and relative change in patient-reported pain intensity

ROC analysis (Fig. 2) was performed to identify meaningful thresholds for both absolute (Δ Pain) and relative (Δ_r Pain) changes in patient-reported pain scores reduction after lumbar infiltration, using 'pain level satisfaction' as a reference variable. The ideal point on an ROC curve⁶⁵, would be in the upper left corner (0, 1), representing the best trade-off between specificity (Sp) and sensitivity (Se) for a diagnostic test (Sp 100%, Se 100%)⁶⁶. In our analysis, the optimal cut-off points on these curves would represent the smallest change in pain score, absolute and relative, that best distinguishes between satisfied and unsatisfied patients. The optimal cut-off point can be determined using several approaches⁶⁷.

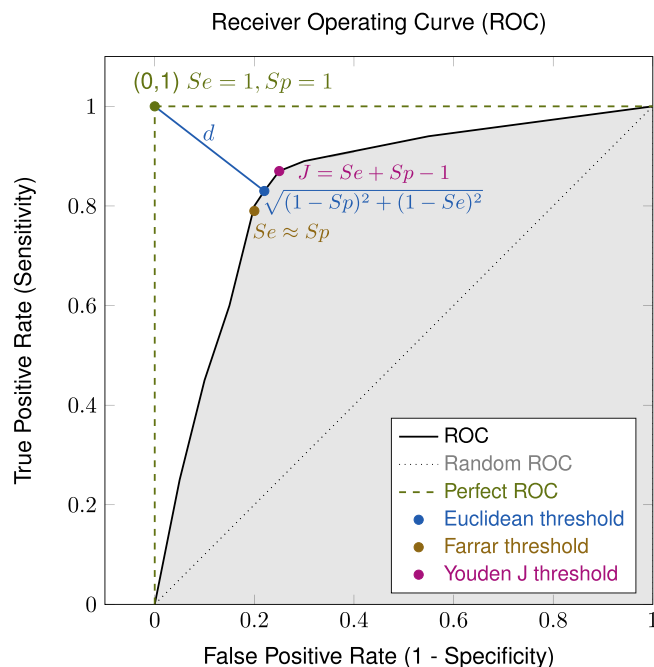


Fig. 2. The receiver operating characteristic (ROC) curve provides an assessment of the predictor's efficacy by plotting true positive rate or sensitivity (Se) against false positive rate, or 1-specificity across multiple binarization thresholds. The ideal ROC curve (dashed green), indicative the ideal classification, and the diagonal (dotted gray) representing random prediction are included for comparative purposes. The highlighted (gray) area under the curve (AUC) serves as a summary statistic for overall classifier performance. The methods employed for determining a threshold that optimally equilibrates sensitivity and specificity comprise: the Euclidean Distance approach (blue), involving the minimization of distance (d) to the ideal threshold at left corner $(0, 1)$; Youden's Index, which focuses on maximizing the disparity between the $Se + Sp - 1$, finding the point furthest from the diagonal ROC; the Farrar Method (bronze), where the Sp equals Se , representing a balance between false positives and the complementary of false negatives.

- The Euclidean distance method: Finds the ROC point closest to the ideal cut-off $(0,1)$ by minimizing the Euclidean distance⁶⁷.
- Youden method: The Youden Index⁶⁸ measures diagnostic performance by summing sensitivity (Se) and specificity (Sp): $J = Se + Sp - 1 = Se - (1 - Sp)$. The optimal cut-off is at the maximum Youden Index, with higher values indicating greater effectiveness by maximizing the vertical distance from the random ROC (diagonal line).
- Farrar method: The Farrar method determines the threshold at which sensitivity and specificity are equal⁶⁷, to equally identify satisfied (Se) and unsatisfied patients (Sp).

Predictor of treatment success based on change in patient-reported pain intensity

To further assess the robustness of our classification approach under a clinically meaningful definition of success, we conducted an additional evaluation using the smallest relative change in baseline pain scores ($\Delta_r \text{Pain}$) as a threshold. Specifically, based on the optimal cut-off identified in our ROC analysis, we stratified participants into "satisfied" or "dissatisfied" groups according to whether their relative reduction in self-reported pain exceeded this threshold at T_1 . This aims to reflect a more clinically relevant perspective of improvement, since percentage-based reductions in pain often better capture individual differences than absolute changes alone.

In this simplified analysis, we applied the same preprocessing and cross-validation schemes but focused exclusively on the set of baseline classifiers without hyperparameter optimization (see previous section). By benchmarking these models, we obtained a clear view of how different algorithms perform when using a clinically significant cut-off for pain relief, rather than the original dichotomous satisfaction variable.

Results

Demographics and baseline pain characteristics

Table 1 presents the sociodemographic characteristics of 212 patients stratified by pain level satisfaction, with 158 patients (75%) in the satisfied group and 54 patients (25%) in the dissatisfied group, with the mean age being 67.8 years ($SD = 2.7$), or gender distribution, with 53% of the patients being female. The analysis did not reveal statistically significant differences in age and gender distribution for the 2 groups. Educational backgrounds were comparable between groups, with vocational apprenticeships being the most common qualification, followed by higher vocational education. Professional status was similarly distributed across groups, with nearly half of


	Satisfied (n=158)	Dissatisfied (n=54)	
Demographics			
Age, mean (SD)	67.8 (±12.7)	63.0 (±16.4)	0.39
Sex—female, N (%)	83 (52.5)	33 (61.6)	0.72
Worktime			
Unknown, N (%)	116 (73.4)	33 (61.1)	0.43
Full time, N (%)	25 (15.8)	13 (24.1)	0.59
Part time, N (%)	17 (10.8)	8 (14.8)	1.00
Education			
Unknown, N (%)	7 (4.4)	2 (3.7)	1.00
No school, N (%)	0 (0.0)	1 (1.9)	1.00
Secondary school, N (%)	9 (5.7)	2 (3.7)	1.00
Vocational apprenticeship, N (%)	58 (36.7)	22 (40.7)	1.00
Vocational-professional baccalaureate, N (%)	9 (5.7)	3 (5.6)	1.00
Academic baccalaureate, N (%)	6 (3.8)	1 (1.9)	1.00
Higher vocational education, N (%)	32 (20.3)	11 (20.4)	1.00
University degree, N (%)	27 (17.1)	9 (16.7)	1.00
Doctorate, N (%)	10 (6.3)	3 (5.6)	1.00
Profession (not exclusive)			
Self-employed, N (%)	28 (17.7)	4 (7.4)	0.42
Student, N (%)	2 (1.3)	1 (1.9)	1.00
Housework, N (%)	32 (20.3)	3 (5.6)	0.33
Retired, N (%)	80 (50.6)	21 (38.9)	0.51
Incapacity, N (%)	9 (5.7)	8 (14.8)	0.39
Unemployed, N (%)	1 (0.6)	0 (0.0)	1.00
Profession: other, N (%)	2 (1.3)	1 (1.9)	1.00

Table 1. Descriptive statistics of sociodemographic variables stratified by reported pain level satisfaction. Data are presented as mean (± standard deviation) for continuous variables and frequency (percentage) for categorical variables. Statistical comparisons p-values are corrected via Benjamini-Hochberg method.

the participants in both groups being retired. No statistically significant differences were observed between the groups in terms of education or employment status.

The majority of patients reported significant baseline pain, with a mean pain level of 5.8 ± 2.3 on the NRS (0-10), indicating substantial discomfort before treatment. Notably, 75% of patients had baseline pain levels of 5 or higher. After treatment, patients experienced a reduction in pain, with the mean pain level decreasing to 3.9 ± 2.6 , reflecting an average pain reduction of 2.7 ± 2.5 points on the NRS. At baseline (T_0), dissatisfied patients reported maximum higher back or leg pain levels compared to satisfied patients, 6.4 ± 2.1 vs. 7.3 ± 2.4 ($p = 4.865e - 03$). After treatment (T_1), this difference became more pronounced, with satisfied patients showing lower pain scores of 3.1 ± 2.2 while dissatisfied patients maintained high pain levels ($6.2 \pm 1.9, p = 5.685e - 17$)

Baseline predictive models of treatment satisfaction

We evaluated baseline classifiers with stratified cross-validation and quantified performance using per-fold mean AP, AUC, accuracy, F1-score, Cohen’s Kappa, and MCC without hyperparameter optimization, summarized in Fig. 3.

The RF demonstrated superior performance in terms of discriminative ability, achieving the highest AP of 0.856 on the full cross-validated test set. In terms of overall metrics, the model showed an accuracy of 70.3%, a precision of 78.1%, and a recall of 83.5%, resulting in an F1-score of 0.807. The MCC and Cohen’s Kappa scores were 0.163 and 0.161, respectively, suggesting a weak correlation and fair agreement between predicted and actual classes. The ROC AUC was 0.731, indicating fair discriminative ability.

Optimized classifier performance

Although the optimal hyperparameters for the RF classifier exhibited slight variations across different folds, certain consistent trends were discernible. The number of estimators demonstrated a wide range, ranging from 21 to 109 trees. Tree depth varied, with some folds using shallow trees (depths of 4–6) and others can grow without a predetermined limit. Feature sampling consistently targeted 70–90% of features. Conservative parameters for leaf nodes and node splitting required minimal samples (2–5 for leaves, 6–10 for nodes).

The tuned RF classifier demonstrated moderate predictive performance across multiple evaluation metrics, as visualized in Fig. 4a. The model achieved a ROC AUC of 0.688, indicating fair discriminative ability. The Precision-Recall curve (right of Fig. 4a) revealed an AP score of 0.846, showcasing a good precision-recall trade-off. The confusion matrix (Fig. 4a) shows the model correctly identified 131 satisfied and 27 dissatisfied patients,

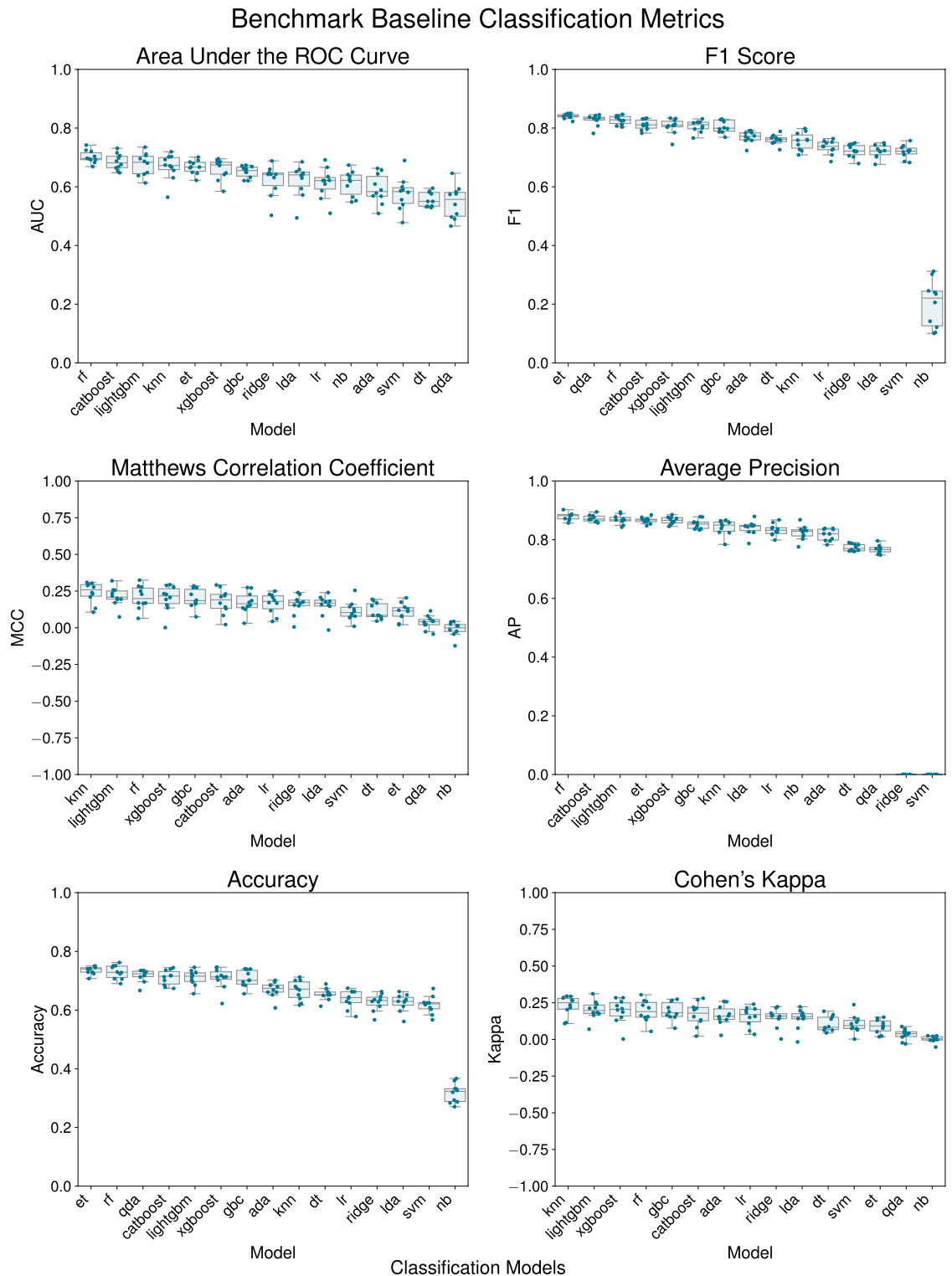


Fig. 3. The boxplots illustrate the distribution of each metric across cross-validation folds for each classifier model, facilitating a performance comparison of baseline classifiers utilizing various metrics: Area Under the ROC Curve (AUC), F1-score, Matthews Correlation Coefficient (MCC), Average Precision (AP), Accuracy, and Cohen's Kappa. The classifiers included K-Nearest Neighbors (KNN), Random Forest (RF), Extra Trees (ET), Gradient Boosting Classifier (GBC), XGBoost, Ridge Classifier (RC), Linear Discriminant Analysis (LDA), LightGBM, Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), AdaBoost, Naive Bayes (NB), and Quadratic Discriminant Analysis (QDA). RF achieved the highest mean Average Precision (AP) of 0.879 ± 0.012 and an Area Under the Curve (AUC) of 0.702 ± 0.020 . RF also exhibited strong classification accuracy, with a mean of 0.729 ± 0.022 and an F1-score of 0.827 ± 0.014 . In contrast, K-Nearest Neighbors (KNN) achieved the highest Cohen's Kappa value of 0.228 ± 0.065 and the highest Matthews Correlation Coefficient (MCC) value of 0.239 ± 0.068 .

while misclassifying 29 as satisfied and 25 as dissatisfied out of 212 patients. This distribution highlights a slight class imbalance, with the model demonstrating higher sensitivity for satisfied patients.

The F1 score of 0.824 reflected balanced performance between precision (0.819) and recall (0.829). Further evaluation of performance metrics provided additional insights: the MCC and Cohen's Kappa values of 0.296 indicated weak correlation and fair agreement between predicted and actual classes. The balanced accuracy of 64.6% further underscored the model's moderate classification capabilities, with an overall accuracy of 73.6%.

Feature importance analysis of optimized classifier

The SHAP analysis of the common features selected across all folds revealed that pain self-efficacy features were the most significant predictors of post-treatment pain level satisfaction, as illustrated in the bar plot in Fig. 4b.

The strongest predictor was patients' perceived ability to stay active despite pain (*pain_self_efficacy_activity*), followed by patient age and days of activity limitation due to back pain (*comi6_socialdisability_T0*). The duration of severe back symptoms (*duration_back_complaints*) and baseline leg pain intensity (*comi2_legpain_T0*) showed

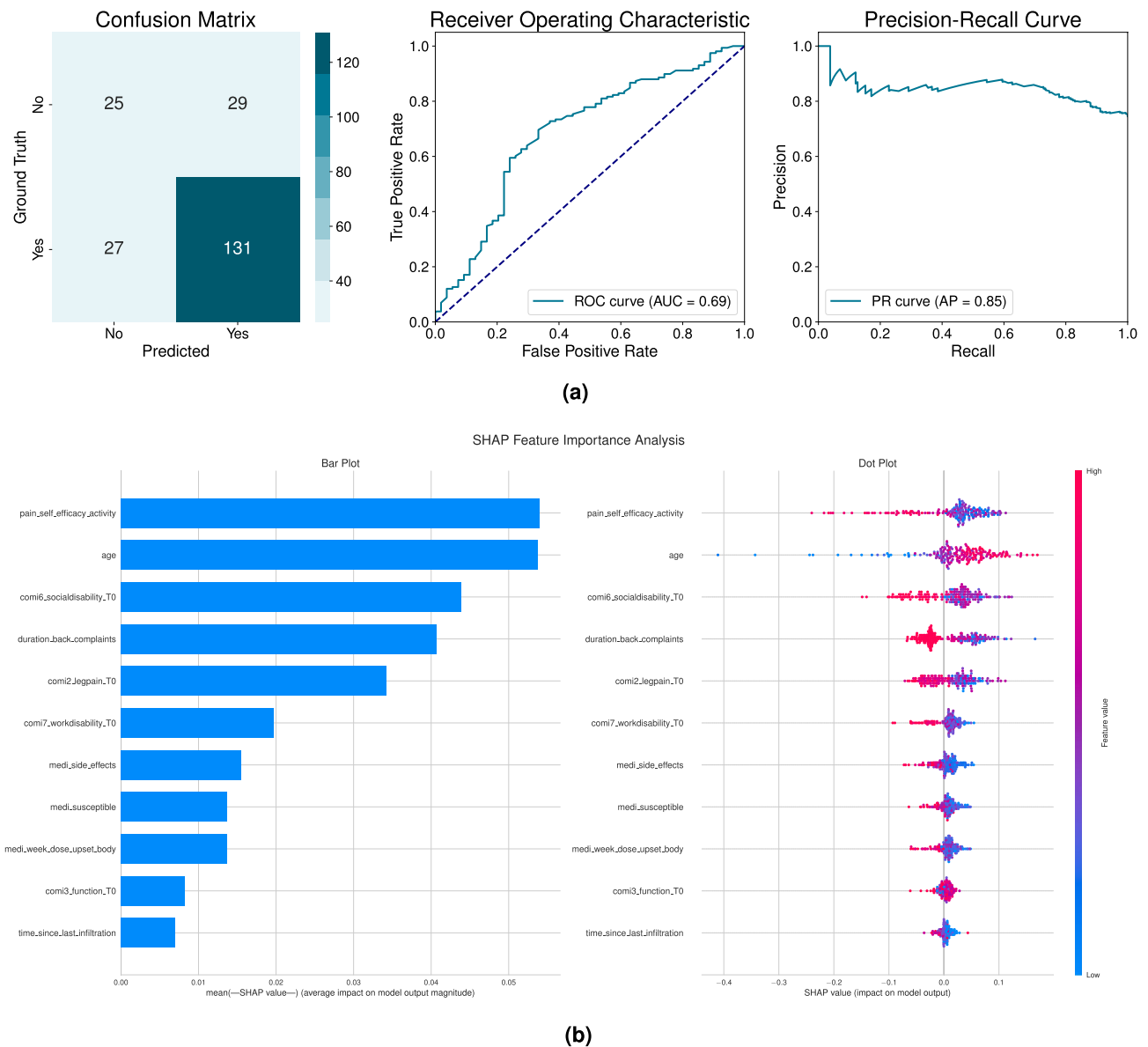


Fig. 4. (a) Classification results of the optimized random forest (RF) model for patient reported pain level satisfaction two weeks after lumbar steroid injection treatment (left) the confusion matrix shows the true and false positives and negatives classification (middle) receiver operating curve (ROC) curve (right) precision-recall curve illustrating the relationship between precision and recall across classification thresholds. (b) shap feature importance analysis plots: bar plot (left) illustrates the average relative significance and dot plots (right) reveals how each feature affects outcomes, where red dots indicate higher values and blue dots show lower values. The bar plot underscores the influence of pain self-efficacy activity, age and disability as well as pain temporal variables in the model outcome.

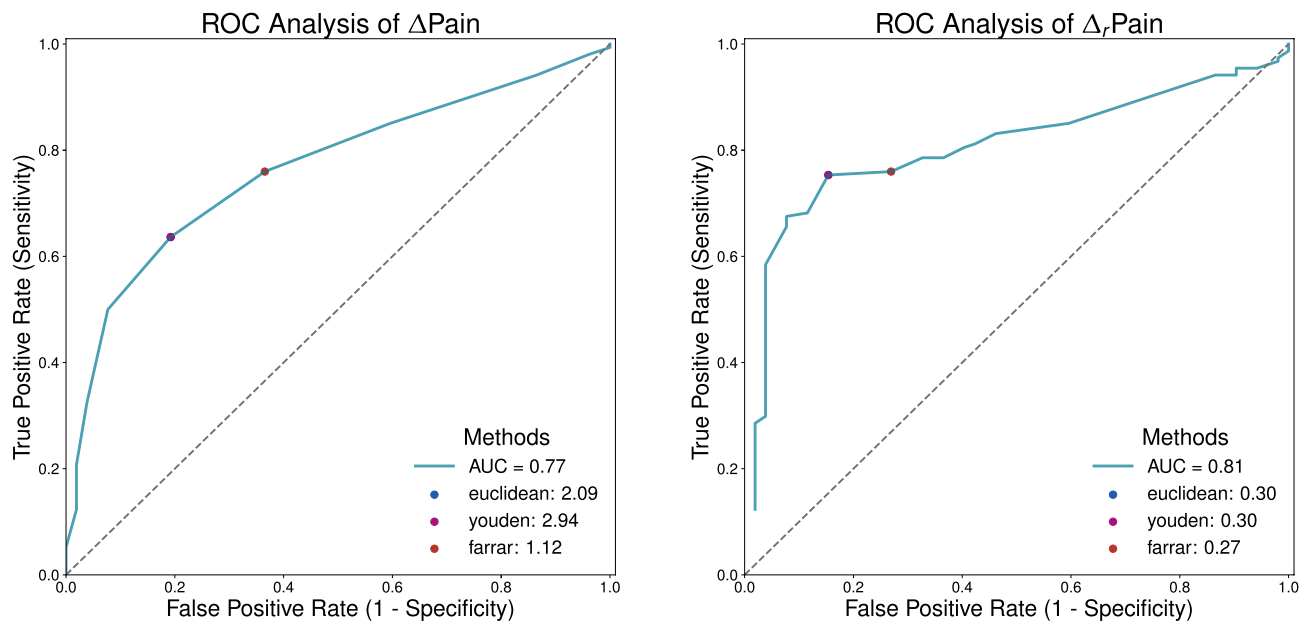


Fig. 5. ROC curve analysis for absolute pain level (ΔPain) and relative ($\Delta_r\text{Pain}$) change in reported maximum pain between baseline and 2-weeks after treatment. The figure illustrates the best thresholds of both changes in pain scores that can be used in distinguishing between satisfied and dissatisfied patients following lumbar steroid injection therapy.

moderate predictive influence. Medication sensitivity factors including history of side effects (*medi_side_effects*), self-reported susceptibility (*medi_susceptible*), and reaction to low doses (*medi_week_dose_upset_body*) exhibited weaker but statistically meaningful associations. Work disability days (*comi7_workdisability_T0*) and baseline functional impairment (*comi3_function_T0*) completed the predictor influence ranking.

Clinical important change in patient-reported pain intensity and outcomes

The results in Fig. 5 present optimal thresholds for absolute and relative changes in pain scores from ROC curve analysis to distinguish between satisfied and dissatisfied patients after steroid injection therapy.

The ΔPain threshold achieved an AUC of 0.77, while the $\Delta_r\text{Pain}$ cut-off point had an AUC of 0.81, indicating a strong potential to discriminate between satisfied and dissatisfied patients.

For the absolute change in pain ΔPain , the Euclidean distance method identified an optimal threshold of 2.09 (sensitivity: 0.64, specificity: 0.81), while the Youden index yielded a higher threshold of 2.94 with identical sensitivity (0.64) and specificity (0.81). The Farrar method suggested a lower threshold of 1.12, achieving higher sensitivity (0.76) but lower specificity (0.63).

For the absolute change in pain (ΔPain), the Euclidean distance method determined a threshold of 2.09, characterized by a sensitivity of 0.64 and a specificity of 0.81. Meanwhile, the Youden method identified a threshold of 2.94 accompanied by a sensitivity of 0.64 and a specificity of 0.81. The Farrar method yielded a threshold of 1.12, with a sensitivity of 0.76 and a specificity of 0.63. Considering the relative change in pain ($\Delta_r\text{Pain}$), both the Euclidean distance and Youden methods determined an optimal cut-off point of 0.30, with a sensitivity of 0.75 and a specificity of 0.85. The Farrar method identified a threshold of 0.27, with a sensitivity of 0.76 and a specificity of 0.73.

Taking into account the consistency between methods and the balance between sensitivity and specificity, a ΔPain of 2.09 and a $\Delta_r\text{Pain}$ of 0.30 were identified as the best thresholds indicating the minimal clinically relevant enhancement in pain relief after steroid injection therapy.

Baseline classifiers for treatment success based on change in patient-reported pain intensity

A re-assessment of the baseline classifiers was conducted to predict patient satisfaction based on the previously established clinically meaningful threshold of 30% relative pain reduction ($\Delta_r\text{Pain} = 0.30$). RF demonstrated superior performance in terms of discriminative ability, achieving the highest mean AP of 0.757 ± 0.020 and AUC of 0.651 ± 0.020 . Extra Trees (ET) exhibited strong performance with a mean AUC of 0.628 ± 0.024 and accuracy of 0.579 ± 0.020 for relative pain reduction. MCC values were consistently low across all models, ranging from 0.003 for QDA to 0.196 for RF indicating weak correlation between predicted and actual classes. Similarly, Cohen's Kappa values were also modest, with the highest value being 0.190 for RF, followed by 0.172 for AdaBoost.

Discussion

Our findings underscore the complex and multifaceted nature of CLBP and highlight the potential of data-driven predictive modeling to refine therapeutic strategies for lumbar steroid injection therapy. The results demonstrate

that baseline characteristics, particularly psychosocial factors such as pain self-efficacy, play a crucial role in determining patient satisfaction after treatment. Previous research showed the importance of expectation on treatment outcome¹⁶. These insights provide a strong foundation for the development of personalized treatment plans aimed at optimizing outcomes in CLBP patients.

The predictive model developed in this study represents a meaningful advance in understanding and addressing the heterogeneity of treatment responses. By highlighting individualized predictors like pain self-efficacy, it enables personalized treatment strategies and better resource allocation. Additionally, the model's use of outcome-driven metrics, such as clinically meaningful pain reduction thresholds, provides actionable insights for standardizing evaluations and optimizing patient-centered care in CLBP. Although our model achieved high average precision (AP: 0.846), the moderate Area Under the Curve (AUC: 0.688) and relatively low Matthews Correlation Coefficient (MCC: 0.296) warrant careful interpretation within the clinical context. This metric pattern reflects the well-documented behavior of evaluation measures in imbalanced datasets rather than fundamental model limitations⁶⁹. The MCC's conservative nature in imbalanced scenarios often produces low values even when overall model performance remains clinically meaningful⁷⁰. From a clinical decision-making perspective, our model's high AP is particularly valuable, as correctly identifying patients likely to benefit from treatment often takes precedence over achieving perfect metric balance. The moderate AUC values still represent fair to good discriminative ability, with values above 0.7 generally considered clinically useful for decision support. Importantly, our decision to preserve the natural class distribution (75% satisfied patients) maintains clinical relevance, as recent evidence suggests that class imbalance corrections can harm model calibration and lead to systematic risk overestimation in clinical prediction models⁷¹.

The prominence of pain self-efficacy measures as key predictors suggests that our model captures clinically meaningful patterns rather than simply exploiting class imbalance. However, future clinical implementation will require careful attention to model calibration to ensure that predicted probabilities accurately reflect true outcome risks for individualized patient stratification.

The feature importance analysis underscores the multidimensional nature of CLBP outcomes, with psychosocial factors emerging as the most significant predictors of patient satisfaction. Specifically, SHAP analysis revealed that *pain self-efficacy activity*—the ability to stay active despite pain—was the strongest contributor to post-treatment satisfaction. This finding highlights the critical role of psychological resilience in shaping patient outcomes, aligning with prior research emphasizing the importance of self-efficacy in chronic pain management⁷². Demographic and functional variables emerged as significant contributors to treatment outcomes in CLBP. Patient age and days of activity limitation due to back pain (*comi6_socialdisability_T0*) ranked as the second and third most influential predictors, respectively. These findings suggest that older patients and those with greater activity limitations may experience worse treatment outcomes, consistent with prior research highlighting the impact of age and functional capacity on recovery trajectories in CLBP⁷³.

Clinical factors, including the duration of severe back symptoms (*duration_back_complaints*) and baseline leg pain intensity (*comi2_legpain_T0*), also demonstrated moderate influence. These variables reflect the complex interplay between symptom severity and treatment response, as supported by studies identifying symptom duration and baseline pain intensity as important prognostic factors for pain reduction and disability improvement in multidisciplinary treatment programs for LBP^{74,75}.

The ROC analysis revealed the minimum score for pain reduction that patients perceive as beneficial after steroid injection therapy. The relative change in pain intensity demonstrated superior discriminative capabilities based on the ROC-AUC analysis (Fig. 5). The AUC for the relative change threshold was slightly higher (0.81 vs 0.78) than that for the absolute change threshold, suggesting that considering the percentage reduction in pain might be more accurate in predicting patient satisfaction than the absolute change. This superiority can be explained by the percentage reduction's ability to account for baseline variability in pain intensity. Patients with higher baseline pain levels may require larger absolute reductions to perceive meaningful relief, while those with lower baseline levels may find smaller absolute reductions sufficient. By normalizing pain reduction relative to the initial intensity, percentage change offers a more individualized and context-sensitive measure, better capturing patient satisfaction across a diverse population. The identified pain reduction thresholds, consisting of an absolute change of 2.0 NRS points and a relative reduction of 30%, are consistent with the range of reductions previously documented in patients with acute pain²⁸. These thresholds provide objective measures for assessing treatment outcomes, managing patient expectations, and standardizing CLBP steroid injection evaluations in clinical settings.

Limitations

Despite its strengths, our study has several limitations. One limitation stems from the secondary analysis design, which constrained our access to comprehensive diagnostic documentation. The original cohort was collected to investigate treatment expectations rather than injection decision-making protocols, resulting in limited systematic diagnostic information typically required for prospective injection studies. Specifically, we lacked detailed records of the diagnostic criteria used to identify facet joint pain, imaging findings (MRI, CT) informing injection site selection, and comprehensive clinical histories guiding level-specific targeting decisions. While the dataset included information on previous lumbar infiltrations, medication use, and pain characteristics, these variables provide only partial insight into the patient's clinical background. Having said this, for all patients included in our study, the injection site was chosen based on a thorough clinical examination and imaging (MRI and/or CT), following a standard clinical procedure at Schulthess Clinic.

Methodological limitations further affect our findings. The relatively small sample size (n=212) may have limited the ability of the model to generalize to broader populations. Furthermore, the use of PROM as the primary endpoint, while clinically relevant, is subjective in nature and may be influenced by recall bias or patient expectations. Furthermore, the dichotomization of the satisfaction outcome variable may oversimplify

the complexity of pain level satisfaction. Although this binary classification is necessary for analysis, it can potentially lead to a loss of granularity with respect to degrees of satisfaction and their contributing predictors. Future models should employ ordinal or continuous outcomes for deeper insights into patient satisfaction and treatment response. Exploring the integration of longitudinal data, including repeated assessments of pain and function, may also provide a more dynamic understanding of treatment responses.

Future research should address these limitations. Prospective studies should validate these findings in larger and more diverse cohorts and incorporate objective measures, such as functional imaging or biomarker analysis, to complement self-reported data. Future work should also aim to enhance the data-driven predictive model by incorporating additional relevant factors, with a focus on multimodal variables (e.g. imaging findings) that better capture the heterogeneity of CLBP. The integration of multimodal variables, including comprehensive imaging data and psychosocial assessments, could significantly improve the model's predictive accuracy and clinical applicability. Finally, examining the utility of identified pain reduction thresholds across distinct CLBP phenotypes may offer valuable insights for optimizing conservative treatment strategies.

Conclusion

In summary, this study underscores the intricate nature of CLBP and the potential of predictive modeling to inform more personalized treatment approaches. Psychosocial factors, particularly pain self-efficacy, emerged as significant contributors to patient satisfaction, reinforcing the need to address psychological dimensions alongside physical interventions. The identified thresholds for pain reduction provide practical benchmarks for evaluating treatment outcomes and standardizing clinical practices. Despite promising predictive performance, the model's limitations highlight the necessity for further refinement and validation with larger, more diverse populations. Future efforts should prioritize integrating multimodal data, such as imaging and comprehensive psychosocial assessments, to enhance the predictive power and clinical utility of these models.

Data availability

Anonymized data used in this study will be made available upon reasonable request to zina-mary.manjaly@kws.ch and in compliance with the General Data Protection Regulation (EU GDPR). We publish all code required to reproduce the presented results in our GitLab repository: <https://gitlab.ethz.ch/BMDSlab/publications/low-back-infiltration-outcome-pain-prediction>

Received: 10 February 2025; Accepted: 7 July 2025

Published online: 29 July 2025

References

1. Ferreira, M. L. 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.* **5**, e316–e329. [https://doi.org/10.1016/S2665-9913\(23\)00098-X](https://doi.org/10.1016/S2665-9913(23)00098-X) (2023).
2. Manchikanti, L. Epidemiology of low back pain. *Pain Phys.* **3**, 167–192 (2000).
3. Meucci, R. D., Fassa, A. G. & Faria, N. M. X. Prevalence of chronic low back pain: Systematic review. *Rev. Saude Publ.* **49** (2015).
4. Eloy, O., Patricia, B.-M. L., Gustavo, R.-C. D. & Rafael, G.-A. Effectiveness of three treatment strategies on quality of life for patients with chronic low back pain: A multidisciplinary approach as key to success. *J. Musculoskelet. Disord. Treat.* (2019).
5. Freburger, J. K. et al. The rising prevalence of chronic low back pain. *Arch. Intern. Med.* **169**, 251–258 (2009).
6. Kent, P. M. & Keating, J. L. The epidemiology of low back pain in primary care. *Chiropract. Osteopathy* **13**. <https://doi.org/10.1186/1746-1340-13-13> (2005).
7. Wu, A. et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: Estimates from the global burden of disease study 2017. *Ann. Transl. Med.* **8**, 299 (2020).
8. Koes, B. W., Van Tulder, M. W. & Thomas, S. Diagnosis and treatment of low back pain. *BMJ* **332**, 1430–1434. <https://doi.org/10.1136/bmj.332.7555.1430> (2006).
9. Li, W. et al. Peripheral and central pathological mechanisms of chronic low back pain: A narrative review. *J. Pain Res.* **14**, 1483–1494 (2021).
10. Kahere, M. & Ginindza, T. The prevalence and psychosocial risk factors of chronic low back pain in KwaZulu-Natal. *Afr. J. Prim. Health Care Fam. Med.* **14**, e1–e8 (2022).
11. Sanzarello, I. et al. Central sensitization in chronic low back pain: A narrative review. *J. Back Musculoskelet. Rehabil.* **29**(4), 625–633 (2016).
12. White, A. P., Arnold, P. M., Norvell, D. C., Ecker, E. & Fehlings, M. G. Pharmacologic management of chronic low back pain. *Spine* **36**, S131–S143. <https://doi.org/10.1097/BRS.0b013e31822f178f> (2011).
13. Urits, I. et al. Low back pain, a comprehensive review: Pathophysiology, diagnosis, and treatment. *Curr. Pain Headache Rep.* **23**, 23. <https://doi.org/10.1007/s11916-019-0757-1> (2019).
14. Staal, J. B., de Bie, R., de Vet, H. C., Hildebrandt, J. & Nelemans, P. Injection therapy for subacute and chronic low-back pain. *Cochrane Database Syst. Rev.* **34**, 49–59. <https://doi.org/10.1002/14651858.CD001824.pub3> (2008).
15. Whynes, D. K., McCahon, R. A., Ravenscroft, A. & Hardman, J. Cost effectiveness of epidural steroid injections to manage chronic lower back pain. *BMC Anesthesiol.* **12**, 26 (2012).
16. Müller-Schrader, M. et al. Individual treatment expectations predict clinical outcome after lumbar injections against low back pain. *Pain* **164**, 132–141. <https://doi.org/10.1097/j.pain.0000000000002674> (2022).
17. Koes, B. W., Scholten, R. J., Mens, J. M. & Bouter, L. M. Efficacy of epidural steroid injections for low-back pain and sciatica: A systematic review of randomized clinical trials. *Pain* **63**, 279–288. [https://doi.org/10.1016/0304-3959\(95\)00124-7](https://doi.org/10.1016/0304-3959(95)00124-7) (1995).
18. Borkan, J. M. & Cherkin, D. C. An agenda for primary care research on low back pain. *Spine* **21**, 2880–2884. <https://doi.org/10.1097/00007632-199612150-00019> (1996).
19. Foster, N. E., Hill, J. C., O'Sullivan, P. & Hancock, M. Stratified models of care. *Best Pract. Res. Clin. Rheumatol.* **27**, 649–661. <https://doi.org/10.1016/j.BERH.2013.10.005> (2013).
20. Abbott, A. Evidence base and future research directions in the management of low back pain. *World J. Orthoped.* **7**(3), 156–61 (2016).
21. Huijnen, I. P., Rusu, A. C., Scholich, S., Meloto, C. B. & Diatchenko, L. Subgrouping of low back pain patients for targeting treatments: Evidence from genetic, psychological, and activity-related behavioral approaches. *Clin. J. Pain* **31** (2015).

22. Pneumáticos, S. G., Chatziioannou, S. N., Hipp, J. A., Moore, W. H. & Esses, S. I. Low back pain: Prediction of short-term outcome of facet joint injection with bone scintigraphy. *Radiology* **238**(2), 693–8 (2006).
23. Pauli, J., Starkweather, A. R. & Robins, J. L. W. Screening tools to predict the development of chronic low back pain: An integrative review of the literature. *Pain Med.* (2018).
24. Boissoneault, J., Mundt, J., Robinson, M. & George, S. Z. Predicting low back pain outcomes: Suggestions for future directions. *J. Orthopaed. Sports Phys. Ther.* **47**, 588–592. <https://doi.org/10.2519/jospt.2017.0607> (2017) (PMID: 28859589).
25. Varrassi, G., Moretti, B., Pace, M. C., Evangelista, P. & Iolascon, G. Common clinical practice for low back pain treatment: A modified Delphi study. *Pain Ther.* **10**, 589–604. <https://doi.org/10.1007/s40122-021-00249-w> (2021).
26. Clement, R. C. et al. A proposed set of metrics for standardized outcome reporting in the management of low back pain. *Acta Orthop.* **86**, 523–533 (2015).
27. Haefeli, M. & Elfering, A. Pain assessment. *Eur. Spine J.* **15**, S17–S24. <https://doi.org/10.1007/s00586-005-1044-x> (2006).
28. Olsen, M. F. et al. Pain relief that matters to patients: Systematic review of empirical studies assessing the minimum clinically important difference in acute pain. *BMC Med.* **15**, 35 (2017).
29. Crosby, R. D., Kolotkin, R. L. & Williams, G. Defining clinically meaningful change in health-related quality of life. *J. Clin. Epidemiol.* **56**, 395–407. [https://doi.org/10.1016/S0895-4356\(03\)00044-1](https://doi.org/10.1016/S0895-4356(03)00044-1) (2003).
30. Mannion, A. F. et al. The quality of spine surgery from the patient's perspective: Part 2. Minimal clinically important difference for improvement and deterioration as measured with the core outcome measures index. *Eur. Spine J.* **18**, 374–379 (2009).
31. Hägg, O., Fritzell, P. & Nordwall, A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur. Spine J.* **12**, 12–20. <https://doi.org/10.1007/s00586-002-0464-0> (2003).
32. Cina, A. et al. Methodological considerations in calculating the minimal clinically important change score for the core outcome measures index (comi): insights from a large single-centre spine surgery registry. *Eur. Spine J.* **33**, 4415–4425. <https://doi.org/10.1007/s00586-024-08537-7> (2024).
33. Suzuki, H. et al. Clinically significant changes in pain along the pain intensity numerical rating scale in patients with chronic low back pain. *PLOS ONE* **15**, 1–16. <https://doi.org/10.1371/journal.pone.0229228> (2020).
34. Horne, R. et al. The perceived sensitivity to medicines (PSM) scale: An evaluation of validity and reliability. *Br. J. Health Psychol.* **18**, 18–30. <https://doi.org/10.1111/j.2044-8287.2012.02071.x> (2013).
35. Neumann, M. et al. Psychometrische evaluation der Deutschen version des Messinstruments „consultation and relational empathy (CARE) am Beispiel von Krebspatienten. *Psychother. Psychosom. Med. Psychol.* **58**, 5–15. <https://doi.org/10.1055/s-2007-970791> (2008).
36. Mannion, A. F. et al. The core outcome measures index in clinical practice. The quality of spine surgery from the patient's perspective. Part 1. *Eur. Spine J.* **18**, 367–373. <https://doi.org/10.1007/s00586-009-0942-8> (2009).
37. Chmielewski, B. & Wilski, M. Psychometric properties of chosen scales evaluating disability in low back pain—Narrative review. *Healthcare* **12**. <https://doi.org/10.3390/healthcare12111139> (2024).
38. Deyo, R. A. et al. Outcome measures for low back pain research: A proposal for standardized use. *Spine* **23**, 2003–2013 (1998).
39. Miekisiak, G. et al. Cross-cultural adaptation and validation of the polish version of the core outcome measures index for low back pain. *Eur. Spine J.* **22**, 995–1001 (2013).
40. Genevay, S. et al. Validity of the French version of the core outcome measures index for low back pain patients: A prospective cohort study. *Eur. Spine J.* **23**, 2097–2104 (2014).
41. Qiao, J. et al. Validation of the simplified Chinese version of the core outcome measures index (COMI). *Eur. Spine J.* **22**, 2821–2826 (2013).
42. Mannion, A. F. et al. The Core Outcome Measures Index (COMI) is a responsive instrument for assessing the outcome of treatment for adult spinal deformity. *Eur. Spine J.* **25**, 2638–2648. <https://doi.org/10.1007/S00586-015-4292-4/TABLES/7> (2016).
43. Roland, M. & Fairbank, J. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. *Spine* **25**, 3115–3124. <https://doi.org/10.1097/00007632-200012150-00006> (2000).
44. Ware, J. E. & Sherbourne, C. D. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med. Care* **30**, 473–83 (1992).
45. Davidson, M., Keating, J. L. & Eyres, S. A low back-specific version of the sf-36 physical functioning scale. *Spine* **29** (2004).
46. Jensen, M. P., Turner, J. A. & Romano, J. M. Self-efficacy and outcome expectancies: Relationship to chronic pain coping strategies and adjustment. *Pain* **44**, 263–269. [https://doi.org/10.1016/0304-3959\(91\)90095-F](https://doi.org/10.1016/0304-3959(91)90095-F) (1991).
47. Ali, M. Pycaret: An open source, low-code machine learning library in python (2024). PyCaret version 3.3.2.
48. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLOS ONE* **14**, e0224365. <https://doi.org/10.1371/journal.pone.0224365> (2019).
49. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
50. Bewick, V., Cheek, L. & Ball, J. Statistics review 14: Logistic regression. *Crit. Care* **9**, 112–118 (2005).
51. Taunk, K., De, S., Verma, S. & Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. <https://doi.org/10.1109/iccs45141.2019.9065747> (IEEE, 2019).
52. Hearst, M., Dumais, S., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **13**, 18–28. <https://doi.org/10.1109/5254.708428> (1998).
53. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
54. Webb, G. I. *Naïve Bayes*, 1–2 (Springer, 2016).
55. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2 (Springer, 2009).
56. Safavian, S. & Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **21**, 660–674. <https://doi.org/10.1109/21.97458> (1991).
57. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42. <https://doi.org/10.1007/s10994-006-6226-1> (2006).
58. Breiman, L. Random forest. *Mach. Learn.* **45**, 5–32. <https://doi.org/10.1023/a:1010933404324> (2001).
59. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting, 23–37 (Springer, 1995).
60. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**. <https://doi.org/10.1214/aos/1013203451> (2001).
61. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 11. KDD '16. 785–794. <https://doi.org/10.1145/2939672.2939785> (ACM, 2016).
62. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I. et al. (eds.) *Advances in Neural Information Processing Systems*. Vol. 30 (Curran Associates, Inc., 2017).
63. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, <https://doi.org/10.1186/s12864-019-6413-7> (2020).
64. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. et al. (eds.) *Advances in Neural Information Processing Systems*. Vol. 30 (Curran Associates, Inc., 2017).

65. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* **4**, 627–635 (2013).
66. Akobeng, A. K. Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatr.* **96**, 644–647. <https://doi.org/10.1111/j.1651-2227.2006.00178.x> (2007).
67. Unal, I. Defining an optimal cut-point value in roc analysis: An alternative approach. *Comput. Math. Methods Med.* **2017**, 3762651. <https://doi.org/10.1155/2017/3762651> (2017).
68. Fluss, R., Faraggi, D. & Reiser, B. Estimation of the Youden index and its associated cutoff point. *Biometric. J.* **47**, 458–472. <https://doi.org/10.1002/bimj.200410135> (2005).
69. Li, J. Area under the roc curve has the most consistent evaluation for binary classification, <https://doi.org/10.1371/journal.pone.0316019> (2024). [arXiv: 2408.10193](https://arxiv.org/abs/2408.10193).
70. Zhu, Q. On the performance of Matthews correlation coefficient (mcc) for imbalanced dataset. *Pattern Recognit. Lett.* **136**, 71–80. <https://doi.org/10.1016/j.patrec.2020.03.030> (2020).
71. Carriero, A. et al. The harms of class imbalance corrections for machine learning based prediction models: A simulation study (2024). [arXiv: 2404.19494](https://arxiv.org/abs/2404.19494).
72. Costa, L. D. C. M., Maher, C. G., McAuley, J. H., Hancock, M. J. & Smeets, R. J. E. M. Self-efficacy is more important than fear of movement in mediating the relationship between pain and disability in chronic low back pain. *Eur. J. Pain* **15**, 213–219 (2011).
73. Zhang, Y.-H. et al. Demographic and clinical characteristics associated with failure of physical therapy in chronic low back pain: a secondary analysis from a randomized controlled trial. *Eur. J. Phys. Rehabil. Med.* **60**. <https://doi.org/10.23736/s1973-9087.24.08033-x> (2024).
74. Adnan, R. et al. Determining predictive outcome factors for a multimodal treatment program in low back pain patients: A retrospective cohort study. *J. Manipul. Physiol. Ther.* **40**, 659–667. <https://doi.org/10.1016/j.jmpt.2017.09.001> (2017).
75. Nicholas K, D., Emily J, G., Alexis M, V., Sergey S, T. & Andrew C, H. Prognostic factors for disability and pain outcomes in patients with axial low back pain undergoing a multidisciplinary spine treatment program. *Int. J. Physiatry* **6**. <https://doi.org/10.23937/2572-4215.1510019> (2020).

Acknowledgements

This project was supported by grant (# 380, Jutzeler, Manjaly) of the Strategic Focus Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain (Swiss Federal Institute of Technology). We sincerely thank the patients who participated in this study, contributing their time and experiences to enhance our understanding of chronic low back pain treatment. We also acknowledge and thank the data collection team at Schulthess clinic, as well as the researchers and clinicians of the initial TREXI study¹⁶. ZMM is grateful for support by the Wilhelm-Schulthess Stiftung. We also gratefully acknowledge the support of the Language Center of UZH and ETH Zurich, with special thanks to Kimberly Lewis for her expert assistance in improving the clarity and quality of this manuscript’s scientific writing. For the development of this work, AI-assisted coding systems such as Copilot and Perplexity were used. During the preparation of this manuscript, the author(s) reviewed the text for grammar correctness and enhanced the content with the assistance of AI generative tools (Writefull and Perplexity). After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication. Figure 1 has been designed using resources from Flaticon.com.

Author contributions

M.M. conceptualized the study, developed the codebase, conducted the experiments, and drafted the manuscript. I.D.S. contributed to data preprocessing, performed statistical analysis, and assisted with data interpretation. A.C. provided support with the ROC analysis and participated in the design of the experiments. R.E. contributed to data collection and offered clinical expertise. F.G. provided clinical expertise and assisted in the design of the experiments. C.R.J. and Z.M.M. supervised the project, offered critical revisions, and ensured the scientific rigor of the study. All authors critically reviewed and approved the final version of the manuscript.

Funding

Open access funding provided by Swiss Federal Institute of Technology Zurich

Declarations

Competing interests

C.R.J. serves as a scientific consultant to Abbvie and Mitsubishi Takeda; however, this role had no influence on the design, conduct, or reporting of this study. All other authors declare no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-10907-0>.

Correspondence and requests for materials should be addressed to M.M.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025