Original Research

Inflammatory activity levels on patients with anti-TNF therapy: most important factors and a decision tree model based on REGISPONSER and RESPONDIA registries

David Castro Corredor (ID), **Luis Ángel Calvo Pascual, Eduardo Collantes-Estévez**
and **Clementina López-Medina** (ID)

Abstract

Background: The effectiveness of anti-tumour necrosis factor (TNF) therapy in spondyloarthritis is traditionally associated with factors such as age, obesity and disease subtypes. However, less-explored aspects, such as mental health, socioeconomic status and work type may also play a crucial role in determining inflammatory activity and therapeutic response.

Objectives: To identify the most significant factors explaining inflammatory activity levels in patients treated with anti-TNF therapy and to develop an interpretable machine-learning model with good performance and minimal overfitting.

Design: This is an observational, cross-sectional and multicentre study with socio-demographical and clinical data extracted from the Registry of Spondyloarthritis of Spanish Rheumatology (REGISPONSER) and Ibero-American Registry of Spondyloarthropathies (RESPONDIA) registries.

Methods: We selected patients receiving anti-TNF therapy and applied five feature selection methods to identify key factors. We evaluated these factors using 182 machine learning models, and, finally, we selected a decision tree model that offered comparable performance with reduced overfitting.

Results: Activity levels appear strongly influenced by quality-of-life indicators, particularly the SF-12 physical and mental components and Ankylosing Spondylitis Quality of Life scores. While factors such as age, weight, years of treatment and age at diagnosis have relevance, they are not necessary to obtain a pruned tree with similar cross-validated mean accuracy.

Conclusion: Recognizing the central role of physical and mental well-being in managing disease activity can lead to better therapeutic strategies for chronic disease management.

Keywords: anti-TNF therapy, cross-validated mean accuracy, machine learning, mutual information, rheumatic diseases

Correspondence to:
David Castro Corredor
Rheumatology
Department, Hospital
General Universitario
Ciudad Real, Obispo Rafael
Torija Street, Ciudad Real
13005, Spain
**d.castrocorredor@gmail.
com**

Luis Ángel Calvo Pascual
Department of Quantitative
Methods, ICADE,
Universidad Pontificia de
Comillas, Madrid, Spain

**Eduardo Collantes-
Estévez**
Department of Medical
and Surgical Sciences,
University of Cordoba,
Maimonides Institute for
Research in Biomedicine
of Cordoba (IMIBIC),
Córdoba, Spain

Clementina López-Medina
Rheumatology
Department, Reina Sofia
University Hospital,
Córdoba, Spain

Department of Medical
and Surgical Sciences,
University of Cordoba,
Maimonides Institute for
Research in Biomedicine
of Cordoba (IMIBIC),
Córdoba, Spain

Introduction

Spondyloarthritis remains a challenging disease, often affecting individuals at a young age and leading to lifelong morbidity, representing a significant burden for both the individuals concerned and society. The introduction of anti-tumour necrosis factor (TNF) therapy two decades ago has proven effective in some patient groups in reducing inflammation and disease symptoms, providing unprecedented clinical benefits and a

viable alternative in cases of failure or adverse effects, as seen with nonsteroidal anti-inflammatory drugs (NSAIDs).[1,2] However, the understanding of how TNF inhibitors affect the immune system in patients is still limited, which is relevant since anti-TNF therapy has been associated with infectious complications.[1]

One of the major challenges in anti-TNF treatment is that approximately half of the patients do not show significant clinical responses,[3] suggesting considerable heterogeneity in treatment response. Therefore, identifying which patients have better inflammatory activity in response to anti-TNF could enable the personalization of treatment strategies.

To ensure a robust analysis with a sufficient patient sample, we merged two databases, REGISPONSER and RESPONDIA, based on European Spondyloarthropathy Study Group (ESSG) criteria. This integration enabled us to conduct a comprehensive feature selection using various methods, including mutual information. While the authors previously applied the mutual information technique in studies,[4,5] we have now enhanced its methodological rigour by incorporating bootstrapping and cross-validation. Additionally, we used other feature selection models, such as random forest and logistic regression, and created a ranking to assess the importance of each variable. To further assess the robustness of the selected variables, we analysed cross-validation mean accuracy and mean ROC AUC across multiple machine learning classification models. We identified a pruned decision tree that achieves performance metrics comparable to the best models while using only three features and a shallow depth, reducing overfitting and improving interpretability.

Materials and methods

Study design
We use the multicentric registries REGISPONSER and RESPONDIA as the database, including participants considering 388 patients treated with anti-TNF.

REGISPONSER is a national and multicentre registry that incorporated consecutive SpA patients who fulfilled the ESSG[6] criteria for SpA between March 2004 and March 2007. Thus, patients could have a diagnosis according to their rheumatologist of ankylosing spondylitis (AS), psoriatic arthritis (PsA), inflammatory bowel disease-SpA (IBD-SpA), ReA, u-SpA or Juv-SpA. The study was conducted by the Spanish Group for the Study of Spondyloarthritis of the Spanish Rheumatology Society with 31 participating centres. More information about the design, sampling, recruitment of patients and exclusion and inclusion criteria is detailed in a previous publication by Collantes et al.[7]

RESPONDIA has a similar design and shares the case report form and all of the variables studied with REGISPONSER.[8] It was conducted between 2006 and 2007. Thirty-three centres from eight Latin American countries participated in this registry. The inclusion criteria were the same as in REGISPONSER. Consecutive patients with SpA according to the criteria of the ESSG were included.

The reporting of this study conforms to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.[9]

Variables
A total of 60 variables were analysed in this study (Tables 1 and 2) categorized as follows:

Physical features were recorded, including mean weight and height. Thoracic mobility was assessed through chest expansion, a critical measure for conditions like AS. Additionally, Schober's test was performed as a functional assessment tool.

Patients were stratified by spondyloarthritis (SpA) subtypes, including AS, PsA, and AS+Pso (ankylosing spondylitis with psoriasis). The definition of AS+Pso in the study is based on earlier criteria, as the databases used are not recent; therefore, patients were classified based on psoriasis as a characteristic.

Several clinical symptoms were noted. Buttock pain was a commonly reported symptom, while hip arthritis significantly impacted mobility. Other recorded features included dactylitis, enthesopathy, recurrent tarsitis, and sacroiliitis.

A family history of spondyloarthritis (SpA) was documented, alongside key comorbidities, including inflammatory bowel disease (IBD) and psoriasis. Ocular manifestations such as iritis or uveitis were also reported.

Table 1. Descriptive analysis of numerical variables.

Variables	Mean (SD)	First quartile (25%)	Third quartile (75%)	Missing (%)
Age	44.65 (12.58)	36	54	65 (16.80)
Diagnosis delay	5.15 (7.45)	0	8	54 (13.95)
Duration (years)	16.08 (11.26)	7	22	75 (19.38)
Diagnosis age	33.96 (13.19)	24	43	52 (13.44)
Weight	74.84 (15.18)	65	82	33 (8.53)
Height	166.58 (8.86)	160	172	61 (15.76)
Chest expansion	3.62 (1.99)	2	5	19 (4.91)
Schober's test	3.39 (1.98)	2	4.5	36 (9.30)
Erythrocyte Sedimentation Rate (ESR)	18.88 (17.51)	7	24	41 (10.59)
C-Reactive Protein (CRP)	8.02 (13.00)	1	9	72 (18.60)
ASQoL	6.35 (5.28)	2	10.58	71 (18.35)
SF-12 physical component	34.65 (11.46)	30	42	62 (16.02)
SF-12 mental component	46.90 (13.07)	45	54	62 (16.02)
BASDAI	3.73 (2.48)	1.58	5.51	11 (2.84)
BASFI	3.82 (2.83)	1.3	6.2	14 (3.62)
ASDAS	2.08 (1.20)	1.15	2.81	74 (19.12)

ASDAS, Ankylosing Spondylitis Disease Activity Score; ASQoL, Ankylosing Spondylitis Quality of Life; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index.

Data on treatment approaches were also collected. NSAIDs were commonly prescribed for pain management. The use of corticosteroids, methotrexate, and sulfasalazine was documented as part of the anti-inflammatory and immunosuppressive regimens. The study also captured data on biological treatments, specifically infliximab, etanercept, and adalimumab. Prior treatment with these biologics, as well as corticosteroids, methotrexate, and sulfasalazine, was also documented, allowing for an analysis of both past and present pharmacological strategies and their efficacy.

To assess the impact of disease on function and quality of life, several scales were used. The Ankylosing Spondylitis Quality of Life (ASQoL) score provided an overall assessment of the disease's effect on daily living. The SF-12 Physical and Mental Component Scores were also used to evaluate both physical and mental health, offering a comprehensive view of patient well-being.

Additionally, data on educational level and life conditions were gathered, recognizing the socioeconomic factors that can influence health outcomes. The study also captured disability status, providing insight into how the disease affects patients' abilities to perform daily activities and maintain independence. Patient engagement in physical activity, an important aspect of managing AS, was tracked to understand its role in mitigating disease progression and improving quality of life.

In our study, the target variable was disease activity levels. Since Ankylosing Spondylitis Disease Activity Score (ASDAS) was not available for all patients, we followed a classification criterion previously used in the literature,[10,11] based on the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and the Bath Ankylosing Spondylitis Functional Index (BASFI), incorporating the ASDAS when available. Patients were classified into three categories: "High" activity,

Table 2. Descriptive analysis of categorical variables.

Variable	Categories	Count (%)	Missing (%)
Country	ESP	222 (57.36)	0 (0)
	BRAS	61 (15.76)	
	ARG	45 (11.63)	
	MEX	18 (4.65)	
	VEN	18 (4.65)	
	PORTUGAL	10 (2.58)	
	CHIL	6 (1.55)	
	CR	4 (1.03)	
	PERU	3 (0.78)	
Disability	No	237 (72.92)	62 (16.02)
	Permanent	65 (20.00)	
	Temporary	23 (7.08)	
Gender	Male	273 (70.54)	0 (0)
	Female	114 (29.46)	
Race	White	327 (84.50)	0 (0)
	White-Indigenous	29 (7.49)	
	White-black	17 (4.39)	
	other	7 (1.81)	
	Black-indigenous	3 (0.78)	
	Indigenous-yellow	3 (0.78)	
	Black	1 (0.26)	
Age group	30–49 Years	167 (51.86)	65 (16.79)
	50–69 Years	106 (32.92)	
	10–29 Years	40 (12.42)	
	70+ Years	9 (2.80)	
Life conditions	Good	155 (47.99)	64 (16.53)
	Optimal without luxury	134 (41.49)	
	Deficient	18 (5.57)	
	Inadequate	8 (2.48)	
	Optimal with luxury	8 (2.48)	
Work type	Skilled worker	77 (23.99)	66 (17.05)
	Unskilled worker	72 (22.43)	

(Continued)

Table 2. (Continued)

Variable	Categories	Count (%)	Missing (%)
	Employee	71 (22.12)	
	Intellectual	51 (15.89)	
	Technician	50 (15.58)	
Educational level	University	118 (30.57)	1 (0.25)
	Primary school	116 (30.05)	
	High-school	76 (19.69)	
	Secondary school	69 (17.88)	
	Illiterate	7 (1.81)	
HLA-B27	Positive	200 (77.22)	128 (33.07)
	Negative	59 (22.78)	
Low back pain	Yes	256 (66.15)	0 (0)
	No	131 (33.85)	
Sacroiliac syndrome	No	223 (57.62)	0 (0)
	Yes	164 (42.38)	
Neck pain	No	286 (73.90)	0 (0)
	Yes	101 (26.10)	
Hip arthritis	No	337 (87.08)	0 (0)
	Yes	50 (12.92)	
Lower limb arthritis	No	216 (55.81)	0 (0)
	Yes	171 (44.19)	
Upper limb arthritis	No	304 (78.55)	0 (0)
	Yes	83 (21.45)	
Dactylitis	No	350 (90.44)	0 (0)
	Yes	37 (9.56)	
Recurrent tarsitis	No	304 (93.25)	61 (15.76)
	Yes	22 (6.75)	
Family history of SpA	No	300 (84.27)	31 (8.01)
	Yes	56 (15.73)	
NSAIDs	Habitually	244 (64.21)	7 (1.8)
	On demand	107 (28.16)	
	No	29 (7.63)	
Corticosteroids	No	190 (50.94)	14 (3.61)
	Yes	183 (49.06)	

(Continued)

Table 2. (Continued)

Variable	Categories	Count (%)	Missing (%)
Methotrexate	Yes	232 (61.87)	12 (3.1)
	No	143 (38.13)	
Sulfasalazine	No	211 (57.18)	18 (4.65)
	Yes	158 (42.82)	
Type	AS	222 (57.96)	4 (1.03)
	PsA	47 (12.27)	
	AS-Pso	43 (11.23)	
	USpA	23 (6.01)	
	JSpA	19 (4.96)	
	AS-IBD	18 (4.70)	
	SpA-IBD	6 (1.57)	
	Are	5 (1.31)	
Form	Mixed	208 (55.47)	12 (3.1)
	Axial	126 (33.60)	
	Peripheral	39 (10.40)	
	Entesítica	2 (0.53)	
Axial pain	Yes	338 (87.79)	2 (0.51)
	No	47 (12.21)	
Synovitis	Yes	246 (64.06)	3 (077)
	No	138 (35.94)	
Psoriasis	No	293 (76.70)	5 (1.29)
	Yes	89 (23.30)	
Inflammatory bowel disease	No	353 (92.41)	5 (1.29)
	Yes	29 (7.59)	
Urethritis, cervicitis, and diarrhoea	No	364 (96.30)	9 (2.32)
	Yes	14 (3.70)	
Buttock pain	Yes	232 (61.70)	11 (2.84)
	No	144 (38.30)	
Enthesopathy	No	200 (53.19)	11 (2.84)
	Yes	176 (46.81)	
Sacroiliitis	Yes	318 (83.68)	7 (1.8)
	No	62 (16.32)	

(Continued)

Table 2. (Continued)

Variable	Categories	Count (%)	Missing (%)
Exercise	No	224 (57.88)	0 (0)
	Yes	163 (42.12)	
Iritis/uveitis	No	312 (82.11)	7 (1.8)
	Yes	68 (17.89)	
Dactylitis (peripheral)	No	318 (84.13)	9 (2.32)
	Yes	60 (15.87)	
Previous corticosteroids	No	254 (79.38)	67 (17.31)
	Yes	66 (20.63)	
Previous methotrexate	No	201 (62.62)	66 (17.05)
	Yes	120 (37.38)	
Previous sulfasalazine	No	283 (89.27)	70 (18.08)
	Yes	34 (10.73)	
Previous infliximab	No	173 (55.63)	76 (19.63)
	Yes	138 (44.37)	
Previous etanercept	No	224 (71.79)	75 (19.37)
	Yes	88 (28.21)	
Previous adalimumab	No	263 (86.51)	83 (21.44)
	Yes	41 (13.49)	
Infliximab	Yes	254 (68.83)	18 (4.65)
	No	115 (31.17)	
Etanercept	No	225 (63.38)	30 (8.26)
	Yes	130 (36.62)	
Adalimumab	No	289 (85.76)	50 (12.91)
	Yes	48 (14.24)	
Activity	Low	157 (40.57)	0 (0)
	Medium	132 (34.10)	
	High	98 (25.32)	

AS, ankylosing spondylitis; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; IBD, inflammatory bowel disease.

defined as either ASDAS >3.5 or both BASDAI >4 and BASFI >4; "Low" activity, defined as BASDAI <4 and ASDAS <2.1 and "Medium" activity for all other cases.

Data analysis

We performed data preprocessing to ensure data quality before performing any analyses. This involved merging the two datasets, REGISPONSER and
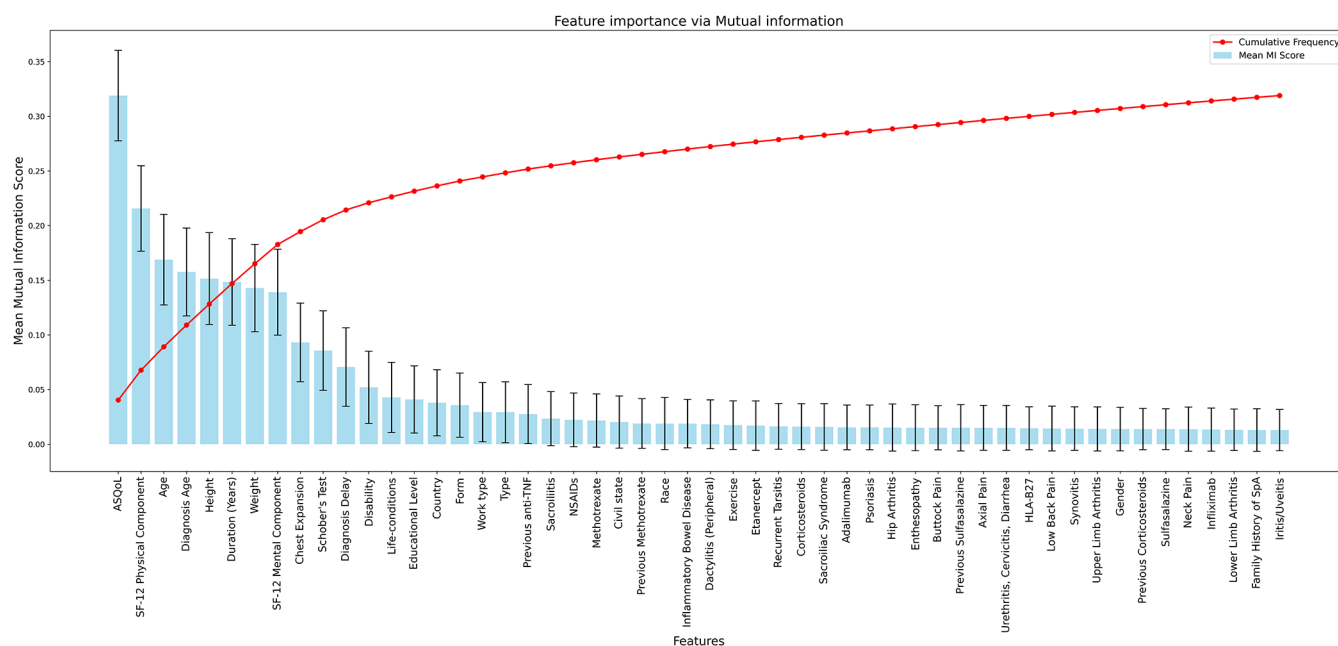
Figure 1. Mutual information test to compute the most relevant variables explaining the activity.

RESPONDIA, while ensuring consistency in columns and categorical variables across both. We examined the dataset for outliers and addressed any extreme values that could potentially distort the model results. Missing values were imputed using the mean of the corresponding columns, a common method that maintains the overall distribution of the data without introducing bias. Then, we made a descriptive analysis of the numerical and categorical variables to understand the features of our study population (Tables 1 and 2).

We employed five distinct feature selection methods to determine the most important variables in assessing activity levels: Mutual Information, Random Forest Feature Importance, Logistic Regression Coefficients, Linear Support Vector Classifier (Linear SVC) and XGBoost Feature Importance. These particular techniques were chosen for their complementary strengths and ability to capture different aspects of feature relevance. Specifically, Mutual Information is adept at identifying non-linear relationships, making it suitable for uncovering complex dependencies (Figure 1). Random Forest Feature Importance leverages ensemble learning to assess the importance of features based on their contribution to reducing impurity across multiple decision trees, effectively handling feature interactions and correlations (Figure 2). By contrast, Logistic Regression Coefficients and Linear SVC focus on

linear relationships, providing straightforward interpretability and understanding of the discriminative power of each feature in classification tasks. Finally, XGBoost Feature Importance employs gradient boosting to evaluate feature significance based on its impact on the loss function of the model.

We recorded accuracy scores and feature importance metrics for each fold, enabling the calculation of mean and standard deviation values. The main results, including mean feature importance, accuracy scores and visualizations for all five models, are provided in the Supplemental Material. To ensure comparability among the different methods, we normalized the importance scores of each method and computed a mean importance score for each feature across all methods. This approach allowed us to rank the features based on their overall significance (Figure 3 and Supplemental Material, Table: results_robust).

Since our dataset was imbalanced in terms of gender, race, and country, we performed a sensitivity analysis by repeating the feature selection process using bootstrap. At each step, we adjusted the dataset to ensure equal representation of males and females, as well as balanced representation across the three most represented countries and racial groups. Additionally, previous studies
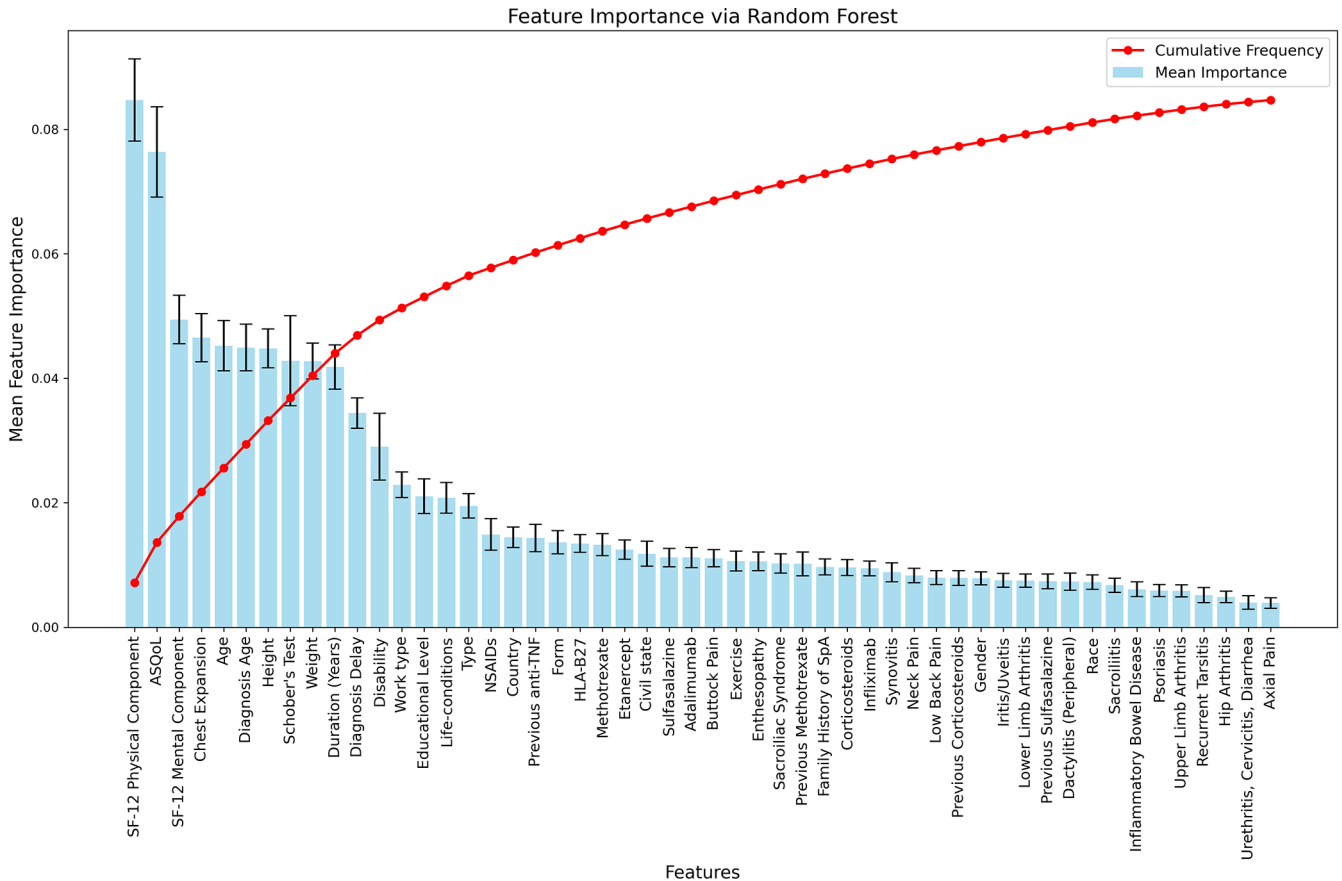
Figure 2. Random forest classifier with Bayesian optimization to compute the most relevant variables explaining the activity.

suggest that the mental component has a greater impact on patients with axial disease compared to those with peripheral disease.[12–15] Therefore, we repeated the feature selection analysis excluding patients with axial disease. We also conducted hypothesis testing and generated boxplots to assess whether the axial-peripheral factor influences the mental component. The codes, tables, and figures for these analyses are available in the Supplemental Material (gender_sens, race_sens, country_sens, and axial_peripheral_sens).

To assess the robustness of the selected features, we evaluated 182 machine-learning models that encompass nearly every standard classifier available in scikit-learn. Our evaluation included linear models (e.g. Logistic Regression, Ridge Classifier), ensemble methods (e.g. Random Forest, Gradient Boosting, AdaBoost), kernel-based methods (e.g. Support Vector Machines with various kernels), instance-based learners (e.g. K-Nearest Neighbors, Nearest Centroid), neural networks (e.g. Multi-Layer Perceptron with different architectures) and

other classifiers such as various Naïve Bayes variants and discriminant analysis techniques. In addition, we incorporated meta-estimators (OneVsRestClassifier, OneVsOneClassifier, StackingClassifier, VotingClassifier), imbalanced data models (e.g. BalancedRandomForestClassifier) and semi-supervised classifiers (LabelPropagation, LabelSpreading, SelfTrainingClassifier). By exploring a broad set of hyperparameter configurations, such as different regularization strengths, tree depths, numbers of estimators, kernel functions and learning rates, and applying five-fold stratified cross-validation for each model, we ensured robust performance estimation while mitigating overfitting. The stratification preserves the original class distribution in every fold, leading to more stable and reliable performance metrics. Together, these choices provide a comprehensive perspective on the effectiveness of the selected features across various classifier types.

For each model, we systematically varied the number of top-ranked features from 3 to 20 and
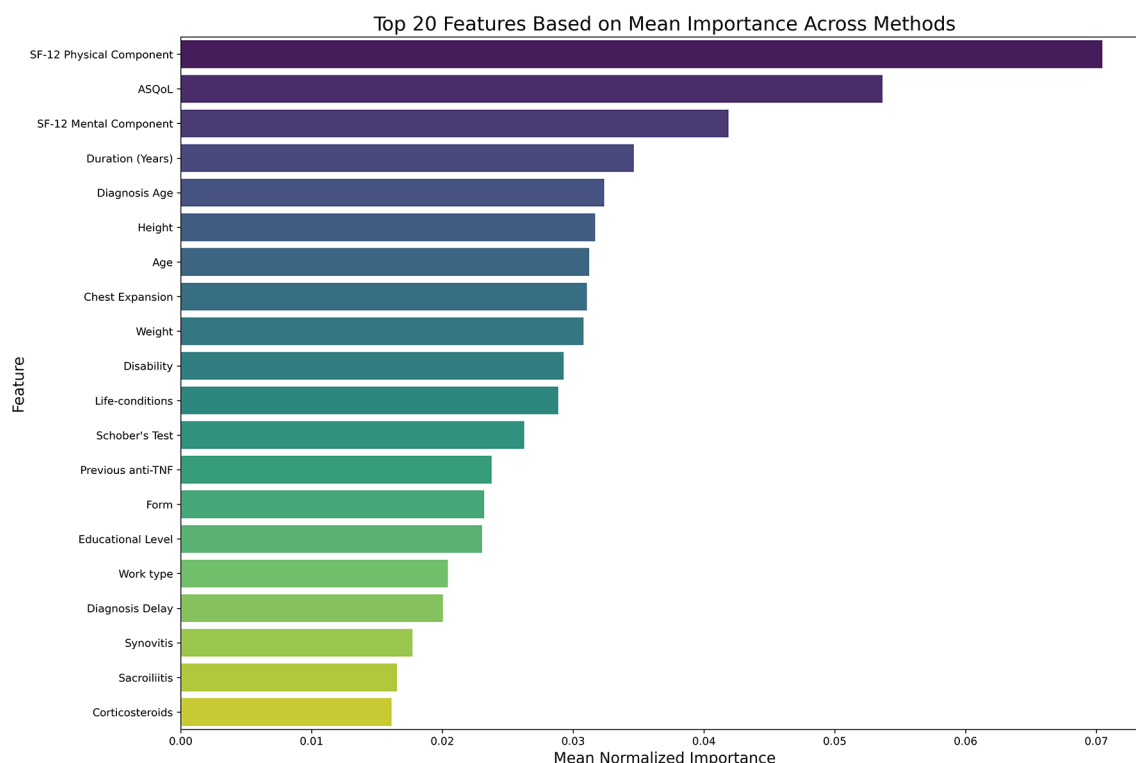
Figure 3. Top 20 features based on mean importance across feature selection methods.

evaluated several hyperparameter combinations, resulting in a total of 3032 model configurations. For every configuration, we collected some performance metrics, which are detailed in the Supplemental Material (Table: model_evaluation). We then ranked the models by sorting them first according to their cross-validated mean accuracy and by the cross-validated mean ROC AUC as a secondary criterion (Table 3). We selected these two metrics because they offer complementary perspectives on model performance while helping to mitigate overfitting. Mean accuracy provides a measure of the proportion of correct predictions, and when averaged over multiple cross-validation folds, it reduces the impact of any single data split. On the other hand, mean ROC AUC measures the ability to discriminate between classes across various thresholds. By integrating both metrics, we ensure that the selected models achieve high accuracy and robust class discrimination, both of which are important for generalizing to unseen data.

The best-performing models (Table 3), in terms of cross-validated mean accuracy and ROC AUC, used more than 11 features, whereas the decision tree of Table 3 used only three. A model that relies on a large number of features is more likely to capture noise and spurious relationships in the training data. Additionally, the three selected features consistently emerged as the most important variables across the sensitivity analyses. Therefore, we opted to use a pruned decision tree with three features, but we changed the maximum depth to three rather than five. Although the deeper tree exhibited marginally better performance metrics, the pruned tree offers better interpretability and further reduces the risk of overfitting by limiting complexity. This trade-off, favouring simplicity and explainability over a slight performance gain, supports the selection of the pruned tree model (Figure 5).

Results
A total of 60 variables were analysed in this study (Tables 1 and 2) categorized as follows:

Demographic variables: 75% of participants were under the age of 54. The cohort had a male predominance, with 70% of patients being men. The majority of patients were from Spain (57.36%),

Table 3. Evaluation of machine learning models hyperparameters, number of features, and cross-validated mean accuracy and ROC AUC, ordered by accuracy from highest to lowest.

Model	Features	Cross-Validates mean accuracy	Cross-Validated mean ROC AUC
Gradient Boosting (learning_rate=0.1) (max_depth=3) (n_estimators=100) (subsample=1.0)	15	0.633011414	0.749499634
Gradient Boosting (learning_rate=0.05) (max_depth=3) (n_estimators=100) (subsample=1.0)	12	0.624100088	0.759358808
Gradient Boosting (learning_rate=0.05) (max_depth=3) (n_estimators=100) (subsample=1.0)	14	0.621290606	0.760120601
Gradient Boosting (learning_rate=0.05) (max_depth=3) (n_estimators=100) (subsample=1.0)	15	0.621158911	0.753480747
Extra Trees (max. features=None)	14	0.618481124	0.738949519
Gradient Boosting (learning_rate=0.05) (max_depth=3) (n_estimators=100) (subsample=1.0)	13	0.618217735	0.753471662
Gradient Boosting (learning_rate=0.05) (max_depth=3) (n_estimators=100) (subsample=1.0)	11	0.615320457	0.744612727
Decision Tree (max_depth=5) (min. samples split=2)	3	0.615320457	0.738570722
Gradient Boosting (LR=. . .) (learning_rate=0.05) (max_depth=3) (n_estimators=100) (subsample=1.0)	18	0.61523266	0.739207242
Gradient Boosting (LR=. . .) (learning_rate=0.1) (max_depth=3) (n_estimators=100) (subsample=1.0)	20	0.615188762	0.733583911

The decision tree highlighted in red achieved similar metrics using only three features.

followed by Brazil (15.76%) and Argentina (11.63%). The racial distribution was predominantly white (84.5%).

Physical features include mean weight (74.84 kg, SD = 15.18 kg) and mean height (166.58 cm, SD = 8.86 cm). Chest expansion was measured at a mean of 3.62 cm (SD = 1.99 cm). Schober's test showed a mean result of 3.39 cm (SD = 1.98 cm).

Regarding the clinical profile, one key indicator was the delay in diagnosis. On average, patients waited 5.15 years (SD = 7.45) before receiving a diagnosis. While 25% were diagnosed within 1 year of symptom onset, some experienced delays of up to 8 years. Additionally, the disease lasted an average of 16.08 years (SD = 11.26), and the mean age at diagnosis was 33.96 years (SD = 13.19).

Disease activity was classified according to the previously described criteria, using BASDAI, BASFI and ASDAS scores. Based on this classification, 157 patients (40%) had "Low" activity, 132 patients (34%) had "Medium" and 98 patients (25%) had "High" activity.

The results of the feature selection process for predicting disease activity revealed that the most influential variables were the SF-12 Physical Component, ASQoL and the SF-12 mental

Decision Tree - Max Depth=3
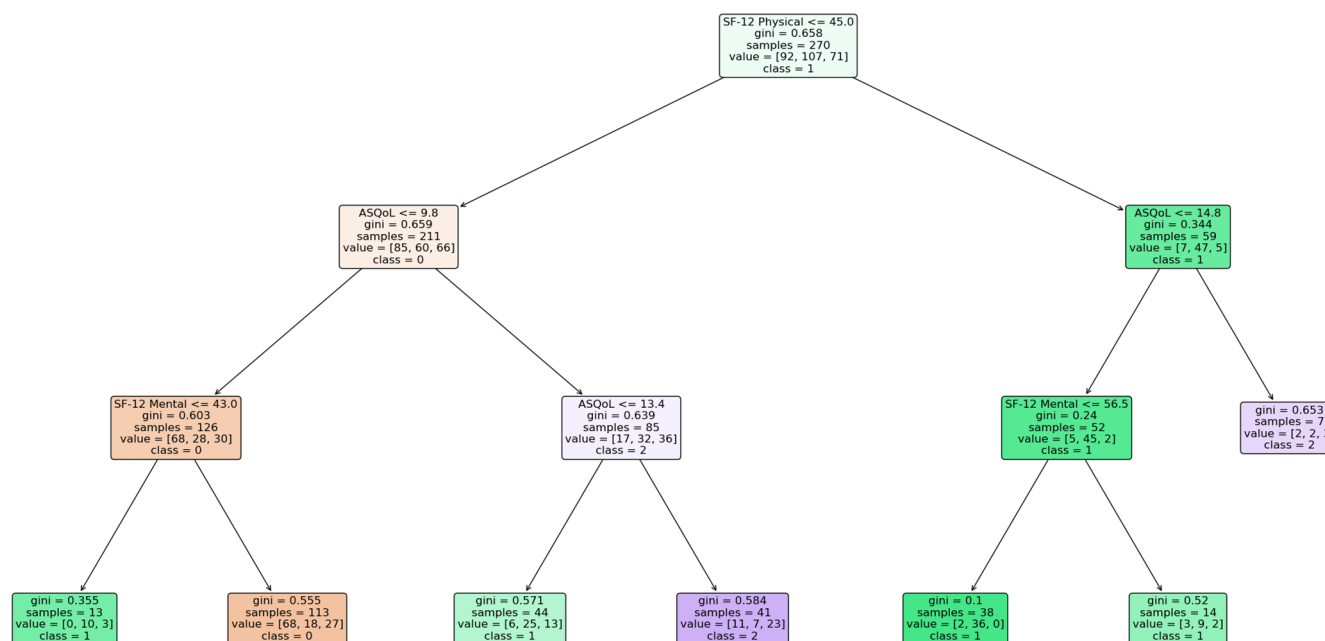
Figure 4. Pruned decision tree explaining activity levels in patients treated with anti-TNF therapy depending on SF-12 physical component, ASQoL and SF-12 mental component.
ASQoL, Ankylosing Spondylitis Quality of Life; TNF, tumour necrosis factor.

component (Figure 3). These variables were used in the tree model presented in Table 3 and in our pruned tree (Figure 4). Other clinical factors, such as height, weight, age, chest expansion, Schober's test and treatment history, including previous use of anti-TNF therapies and corticosteroids, were also found to influence disease outcomes, ranking highly in importance. However, incorporating all of these variables did not substantially improve the performance metrics of the machine learning models (Table 3), suggesting that their contribution to model performance is limited. This indicates that the models are robust even without including these factors. Conversely, our feature selection model identified gender, race, country and axial pain as the least influential features for predicting disease activity.

Our dataset is imbalanced in gender (70.54% men vs. 29.49% women; Table 2), as well as in race (84.50% white) and nationality (57.36% Spanish). The sensitivity analysis using bootstrapping confirmed that the SF-12 mental component, SF-12 physical component and ASQoL consistently rank as the top features, while the other important features like weight, age, chest expansion, Schober's test and treatment history also appear in high-ranking positions, but in varying order (see Supplemental Material: gender_sens, race_sens and country_sens). Additionally, in our dataset, patients with peripheral disease have lower SF-12 mental scores compared to patients with axial disease (U-test, $p = 0.02$). In the sensitivity analysis excluding patients with axial disease, we found that the most important features in the feature selection process remained the SF-12 physical and mental components, as well as ASQoL, with the mental component ranking as the third most important feature (see Supplemental Material: axial_peripheral_sens).

We explain the results of the pruned decision tree (Figure 4). For patients with an SF-12 physical component score below 45, if the ASQoL score is less than 9.8, the SF-12 mental component is used to classify the activity level. A mental component score above 43 indicates a low activity level, whereas a score of 43 or below suggests a medium level. Furthermore, if the ASQoL score is between 9.8 and 13.5, the activity is classified as medium, and if the score exceeds 13.5, the activity is considered high. For patients with an
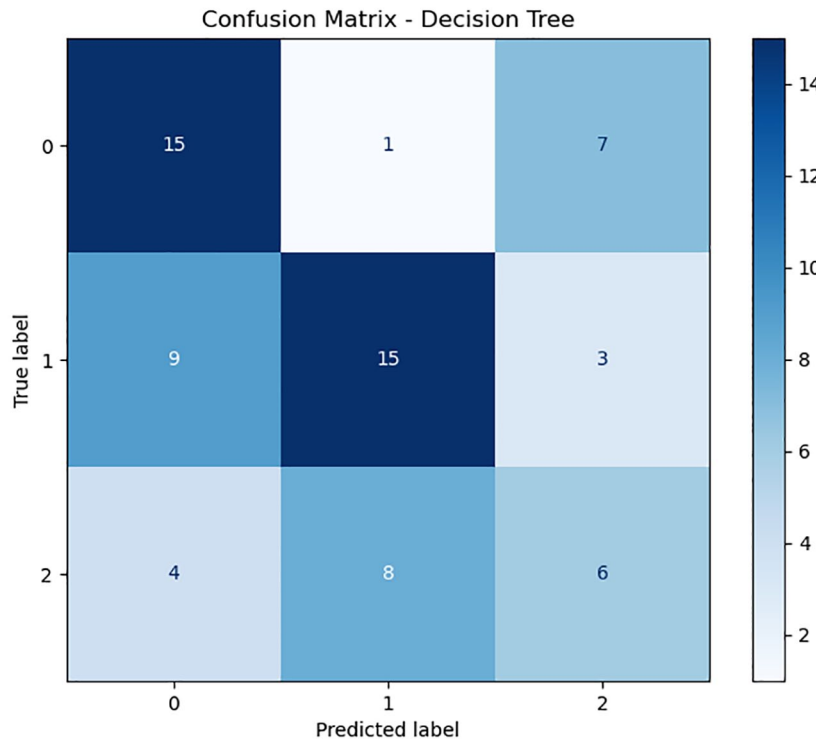
Figure 5. Confusion matrix of the pruned decision tree.

SF-12 physical component score of 45 or above, the ASQoL score is used to differentiate activity levels. Specifically, if the ASQoL score is less than 14.8, the activity is medium, while a score higher than 14.8 indicates high activity. Figure 5 displays the confusion matrix for this tree, and its performance metrics are summarized in the Supplemental Material (see Table: tree). The pruned tree achieved a cross-validated mean accuracy of 0.57 and a mean ROC AUC of 0.70.

Discussion

In the literature, the effectiveness of anti-TNF treatment is influenced by various factors. Advanced age tends to reduce treatment efficacy, particularly in women with axial spondyloarthritis.[16,17] Gender differences are especially notable in older women, who experience lower remission and response rates,[17,18] higher rates of treatment discontinuation[19] and a later onset of symptoms.[20] We noted age and disease duration as important factors (Figure 3); however, in the gender sensitivity analysis, we did not observe significant differences, although, in our sensitivity analysis, we did not differentiate the subgroup of older women.

Obesity further complicates treatment outcomes,[21,22] with studies suggesting a decrease in the BASDAI50 achievement rate from 72.8% in normal-weight patients to 54.5% in overweight patients and 30.4% in obese patients.[23] We observed that weight and height are important factors in explaining disease activity (Figure 3). We agree with these previous studies, as obesity is inherently linked to both weight and height, and these factors likely interact to affect disease progression. The reduction in treatment efficacy may be due to several factors associated with obesity, such as increased systemic inflammation, altered immune responses and the mechanical load obesity places on the body, which can exacerbate musculoskeletal symptoms.

We also noted that the age at diagnosis was an important factor in explaining disease activity (Figure 3). This finding aligns with previous studies.[10,22,23] One possible explanation is that starting treatment early may prevent the progression of damage and inflammation, leading to more effective disease control. By contrast, individuals diagnosed later may have already experienced significant structural damage or prolonged inflammation.

The factors identified in the literature as influencing disease activity in patients with spondyloarthritis, but which we did not find to be significant in our study, include the presence of the HLA-B27 gene[16] and the type of spondyloarthritis. For example, a study by Reveille[24] suggests that radiographic axial SpA patients generally exhibit better responses to anti-TNF treatment compared to those with non-radiographic SpA, particularly when baseline CRP levels are elevated.

In our article, the most important factors for predicting the level of activity were the SF-12 physical and mental components, as well as ASQoL. In Kennedy et al.,[19] it is mentioned that spondyloarthritis is a chronic condition affecting mental health due to persistent pain, functional disability, and uncertainty in treatment. Similarly, in Refs.,[12–15] the authors found that axial patients with symptoms of depression and anxiety had significantly poorer treatment responses and higher rates of treatment discontinuation. Their study also highlighted that ASQoL scores were notably worse in these patients. In our database, peripheral patients exhibited worse mental health compared to axial patients. Furthermore, a sensitivity analysis excluding axial patients revealed that the mental component was the third most important factor in predicting disease activity, suggesting that SF-12 mental health is a crucial feature not only in axial but also in peripheral patients. In this line, a study based on the REGISPONSER cohort found that patients with AS who had two or more comorbidities experienced poorer mental health and lower ASQoL scores.[25]

Acknowledging the importance of physical and mental well-being in disease activity management may help refine therapeutic strategies for chronic conditions. Simple and cost-effective tools like SF-12 and ASQoL questionnaires could be valuable additions to routine clinical practice, offering a practical way to incorporate these factors into patient care.

Now, we discuss some limitations of our study. Machine learning models provide additional insights into the interplay between variables and disease activity. They help us to identify complex, non-linear interactions that traditional statistical methods might overlook; for example, the use of mutual information highlights relationships between ASQoL, age, physical condition and disease activity (Figure 1). Some authors have already employed AI with the same objective as ours.[26,27] However, while machine learning models enhance pattern recognition, they also present challenges regarding the interpretability of their results. Many complex AI models function as 'black boxes', making it difficult to understand the rationale behind their predictions and thereby complicating their clinical applicability. Although our pruned chosen decision tree model offers greater interpretability compared to other methods, it is not without its limitations, as fully understanding its decision logic and clinical implications remains challenging.

One of the primary limitations of our models is the small dataset size (388 patients), which inherently increases the risk of overfitting, especially for complex algorithms like neural networks or boosting methods. To mitigate this, we incorporate cross-validation techniques, specifically Stratified-K-Fold with five-fold, to obtain more reliable estimates of model performance. Additionally, we intentionally select a small number of features (three) based on their importance (Table 3). We also reduced the depth of the tree to avoid overfitting and obtain an interpretable pruned tree.

The cross-validation mean accuracy and mean ROC AUC of the pruned tree (0.57 and 0.70, resp.) are lower than those of the depth-5 tree (0.61 and 0.73) and the gradient boosting model with 15 features (0.63 and 0.74; Table 3). However, we consider this loss in precision acceptable, as it helps prevent overfitting while improving interpretability. In the confusion matrix (Figure 6), the model shows its weakest classification performance in the high activity, frequently misclassifying it as medium activity. A more precise definition of inflammation levels based solely on ASDAS, as recommended by the European Alliance of Associations for Rheumatology (EULAR),[28] could potentially improve classification accuracy. Unfortunately, our database did not consistently include this variable.

We have also found other limitations in our study, such as the fact that it is a cross-sectional observational study that involves the retrospective collection of some data, which limits the ability to conclude causality or temporal relationships between variables. Furthermore, there is a predominance of patients with AS, which may limit the ability to analyse certain associations with other subtypes of spondyloarthritis.
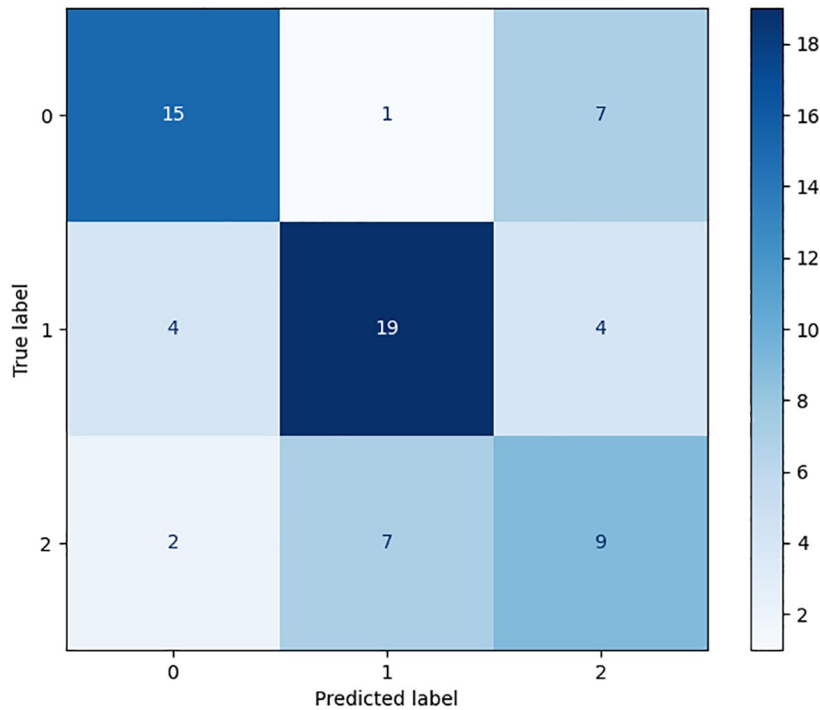
Figure 6. Confusion matrix: decision tree.

Conclusion

We highlight the importance of life conditions and mental health factors, emphasizing the need to integrate quality-of-life measures into therapeutic strategies to improve chronic disease management. Our findings suggest that SF-12 mental and physical components and ASQoL are the most important features for explaining activity levels in patients undergoing anti-TNF therapy. These features appear consistently ranking as the most influential, even across various sensitivity analyses, including those stratified by gender, race, country and non-axial disease type. These questionaries can lead to better therapeutic strategies for chronic disease management. Using these three features, we developed a depth-3 pruned decision tree, which achieved a cross-validated mean accuracy of 0.57 and a cross-validated mean ROC AUC of 0.70 in predicting low, moderate and high inflammation levels. In this tree, we observe that in spondyloarthritis patients with an SF-12 physical component score of ≤ 45 and an ASQoL score of ≤ 9.8, activity levels are medium or low, depending on whether the SF-12 mental component score is above or below 43.

Our findings suggest that age, years of treatment, age at diagnosis, prior treatment, patient height and weight may be important factors, aligning with previous literature. By contrast, variables such as gender, race, the HLA-B27 gene and the type of spondyloarthritis appeared to be less influential in prediction, though their role cannot be entirely ruled out.

Finally, we emphasize again some limitations of our study. First, the relatively small sample size (388 patients) poses a challenge when applying machine learning models, as larger datasets are generally required for more robust predictions. Second, the ASDAS activity variable was not available for all patients, which forced us to rely on alternative metrics such as BASDAI and BASFI to estimate disease activity. This limitation may have affected the precision of our models in capturing inflammation levels. Lastly, it is important to acknowledge that machine learning models are probabilistic, not causal. While our decision tree aims to be as interpretable as possible, it does not establish direct cause-and-effect relationships but rather identifies patterns and associations within the data.

Declarations

Author contributions
David Castro Corredor: Conceptualization; Investigation; Methodology; Project administration; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Luis Ángel Calvo Pascual: Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Writing – original draft; Writing – review & editing.

Eduardo Collantes-Estévez: Funding acquisition; Resources; Validation; Visualization; Writing – original draft; Writing – review & editing.

Clementina López-Medina: Funding acquisition; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Competing interests
The authors declare that there is no conflict of interest.

Availability of data and materials
Data are available on reasonable request. Contact the corresponding author of the manuscript.

Statement on the use of AI for manuscript preparation
During the preparation of this manuscript, the authors utilized an AI-based language model (ChatGPT, OpenAI) to enhance the clarity, style and grammar of the text. The authors independently verified all AI-generated suggestions and maintained full responsibility for the accuracy, integrity and final content of the manuscript, including all interpretations and conclusions.

ORCID iDs
David Castro Corredor https://orcid.org/0000-0001-7315-6274

Clementina López-Medina https://orcid.org/0000-0002-2309-5837

Supplemental material
Supplemental material for this article is available online.

References

1. Menegatti S, Bianchi E and Rogge L. Anti-TNF therapy in spondyloarthritis and related diseases, impact on the immune system and prediction of treatment responses. *Front Immunol* 2019; 10: 382.

2. Pinto AS, Farisogullari B and Machado PM. Predictors of remission in people with axial spondyloarthritis: a systematic literature review. *Semin Arthritis Rheum* 2022; 56: 152078.

3. Wang R, Dasgupta A and Ward MM. Predicting probability of response to tumor necrosis factor inhibitors for individual patients with ankylosing spondylitis. *JAMA Netw Open* 2022; 5(3): e222312.

4. Castro Corredor D and Calvo Pascual LA. Imbalanced machine learning classification models for removal biosimilar drugs and increased activity in patients with rheumatic diseases. *PLoS ONE* 2023; 18(11): e0291891.

5. Calvo Pascual LÁ, Castro Corredor D and Garrido Merchán EC. Machine learning classification of vitamin D levels in spondyloarthritis patients. *Intell Based Med* 2024; 9: 100125.

6. Dougados M, van der Linden S, Juhlin R, et al. The European Spondylarthropathy study group preliminary criteria for the classification of Spondylarthropathy. *Arthritis Rheum* 1991; 34: 1218–1227.

7. Collantes E, Zarco P, Muñoz E, et al. Disease pattern of Spondyloarthropathies in Spain: description of the first national Registry (REGISPONSER) extended report. *Rheumatology (Oxford)* 2007; 46: 1309–1315.

8. Vazquez-Mellado J, Font Ugalde P, Muñoz Gomariz E, et al. Ibero-American Spondylarthritis Registry (RESPONDIA): what is, how came about, who we are and what we do. *Reumatol Clin* 2008; 4: S17–S22.

9. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *PLoS Med* 2007; 4(10): e296.

10. Vastesaeger N, Cruyssen BV, Mulero J, et al.; REGISPONSER Working Group. ASDAS high disease activity versus BASDAI elevation in patients with ankylosing spondylitis as selection criterion for anti-TNF therapy. *Reumatol Clin* 2014; 10(4): 204–209.

11. Machado P, Landewé R, Lie E, et al. Ankylosing Spondylitis Disease Activity Score (ASDAS): defining cut-off values for disease activity states and improvement scores. *Ann Rheum Dis* 2011; 70: 47–53.

12. Garrido-Cumbrera M, Navarro-Compán V, Poddubnyy D, et al. Factors associated with poor mental health in patients with axial spondyloarthritis: results from the International Map of Axial Spondyloarthritis (IMAS). *RMD Open* 2024; 10: e004218.

13. Delcourt C, Fakih O, Prati C, et al. Impact of treatments on fatigue in axial spondyloarthritis: a systematic review and meta-analysis. *Rheumatology* 2025; 64(4): 1585–1597.

14. Zhao SS, Jones GT, Hughes DM, et al. Depression and anxiety symptoms at TNF inhibitor initiation are associated with impaired treatment response in axial spondyloarthritis. *Rheumatology* 2021; 60(12): 5734–5742.

15. Macfarlane GJ, Pathan E, Jones GT, et al. Predicting response to anti-TNFα therapy among patients with axial spondyloarthritis (axSpA): results from BSRBR-AS. *Rheumatology* 2020; 59(9): 2481–2490.

16. Alazmi M, Sari I, Krishnan B, et al. Profiling response to tumor necrosis factor inhibitor treatment in axial spondyloarthritis. *Arthritis Care Res (Hoboken)* 2018; 70(9): 1393–1399.

17. Fernández-Carballido C, Sanchez-Piedra C, Valls R, et al. Female sex, age, and unfavorable response to tumor necrosis factor inhibitors in patients with axial spondyloarthritis: results of Statistical and Artificial Intelligence-Based Data Analyses of a National Multicenter Prospective Registry. *Arthritis Care Res* 2023; 75: 115–124.

18. Albrecht K and Strangfeld A. Gender-specific differences in the diagnosis and treatment of inflammatory rheumatic diseases. *Inn Med (Heidelb)* 2023; 64(8): 744–751.

19. Kennedy J, Kennedy N, Cooksey R, et al. Predicting a diagnosis of ankylosing spondylitis using primary care health records—a machine learning approach. *PLoS One* 2023; 18(3): e0279076.

20. Flouri I, Goutakoli P, Repa A, et al. Distinct long-term disease activity trajectories differentiate early on treatment with etanercept in both rheumatoid arthritis and spondylarthritis patients: a prospective cohort study. *Rheumatol Int* 2024; 44(2): 249–261.

21. Gremese E, Bernardi S, Bonazza S, et al. Body weight, gender and response to TNF-α blockers in axial spondyloarthritis. *Rheumatology* 2014; 53(5): 875–881.

22. Ciurea A, Scherer A, Exer P, et al.; Rheumatologists of the Swiss Clinical Quality Management Program for Axial Spondyloarthritis. Tumor necrosis factor α inhibition in radiographic and nonradiographic axial spondyloarthritis: results from a large observational cohort. *Arthritis Rheum* 2013; 65(12): 3096–3106.

23. Navarro-Compán V, Ermann J and Poddubnyy D. A glance into the future of diagnosis and treatment of spondyloarthritis. *Ther Adv Musculoskelet Dis* 2022; 14: 1759720X221111611.

24. Reveille JD. A registry of ankylosing spondylitis registries and prospects for global interfacing. *Curr Opin Rheumatol* 2013; 25(4): 468–476.

25. Puche-Larrubia MÁ, Ladehesa-Pineda L, Gómez-García I, et al.; REGISPONSER Study Group. Impact of the number of comorbidities on the outcome measures and on the retention rate of the first anti-TNF in patients with Ankylosing Spondylitis. Two-year follow-up in REGISPONSER-AS. *Semin Arthritis Rheum* 2022; 52: 151938.

26. Lee S, Eun Y, Kim H, et al. Machine learning to predict early TNF inhibitor users in patients with ankylosing spondylitis. *Sci Rep* 2020; 10: 20299.

27. Lee S, Kang S, Eun Y, et al. A cluster analysis of patients with axial spondyloarthritis using tumour necrosis factor alpha inhibitors based on clinical characteristics. *Arthritis Res Ther* 2021; 23(1): 284.

28. Ramiro S, Nikiphorou E, Sepriano A, et al. ASAS-EULAR recommendations for the management of axial spondyloarthritis: 2022 update. *Ann Rheum Dis* 2023; 82: 19–34.