Rheumatology, 2023, **62**, 2402–2409
https://doi.org/10.1093/rheumatology/keac645
Advance access publication 23 November 2022
Original article

British Society for Rheumatology RHEUMATOLOGY OXFORD

Clinical science

Machine learning identifies a profile of inadequate responder to methotrexate in rheumatoid arthritis

Julien Duquesne [1,†], Vincent Bouget [1,†], Paul Henry Cournède[2], Bruno Fautrel[3,4], Francis Guillemin[5], Pascal H. P. de Jong[6], Judith W. Heutz[6], Marloes Verstappen[7], Annette H. M. van der Helm-van Mil[7], Xavier Mariette [8,†], Samuel Bitoun [8,†,*]

[1]Scienta Lab, Paris, France
[2]CentraleSupélec, Lab of Mathematics and Computer Science (MICS), Université Paris-Saclay, Gif-sur-Yvette, France
[3]Groupe Hospitalier Pitié Salpêtrière, Service de Rhumatologie, Sorbonne Université – Assistance Publique Hôpitaux de Paris, Paris, France
[4]Inserm UMRS 1136, Équipe PEPITES (Pharmaco-épidémiologie et Évaluation des Soins), Institut Pierre Louis d'Épidémiologie et Santé Publique, Paris, France
[5]APEMAC, Université de Lorraine, Nancy, France
[6]Department of Rheumatology, Erasmus Medical Center, Rotterdam, The Netherlands
[7]Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands
[8]Department of Rheumatology, Université Paris Saclay, INSERM UMR 1184, Hôpital Bicêtre, Assistance Publique-Hôpitaux de Paris, FHU CARE, Le Kremlin Bicêtre, France

*Correspondence to: Samuel Bitoun, Department of Rheumatology, Université Paris Saclay, INSERM UMR 1184, Hôpital Bicêtre, Assistance Publique-Hôpitaux de Paris, FHU CARE, Hôpital Bicêtre 78 avenue du General Leclerc, Le Kremlin Bicêtre France. E-mail: samuel.bitoun@aphp.fr

[†]Julien Duquesne and Vincent Bouget contributed equally to this study. Xavier Mariette and Samuel Bitoun contributed equally to this study.

Abstract

Objectives: Around 30% of patients with RA have an inadequate response to MTX. We aimed to use routine clinical and biological data to build machine learning models predicting EULAR inadequate response to MTX and to identify simple predictive biomarkers.

Methods: Models were trained on RA patients fulfilling the 2010 ACR/EULAR criteria from the ESPOIR and Leiden EAC cohorts to predict the EULAR response at 9 months (± 6 months). Several models were compared on the training set using the AUROC. The best model was evaluated on an external validation cohort (tREACH). The model's predictions were explained using Shapley values to extract a biomarker of inadequate response.

Results: We included 493 therapeutic sequences from ESPOIR, 239 from EAC and 138 from tREACH. The model selected DAS28, Lymphocytes, Creatininemia, Leucocytes, AST, ALT, swollen joint count and corticosteroid co-treatment as predictors. The model reached an AUROC of 0.72 [95% CI (0.63, 0.80)] on the external validation set, where 70% of patients were responders to MTX. Patients predicted as inadequate responders had only 38% [95% CI (20%, 58%)] chance to respond and using the algorithm to decide to initiate MTX would decrease inadequate-response rate from 30% to 23% [95% CI: (17%, 29%)]. A biomarker was identified in patients with moderate or high activity (DAS28 > 3.2): patients with a lymphocyte count superior to 2000 cells/mm^3 are significantly less likely to respond.

Conclusion: Our study highlights the usefulness of machine learning in unveiling subgroups of inadequate responders to MTX to guide new therapeutic strategies. Further work is needed to validate this approach.

Keywords: RA, MTX, treatment response, machine learning, biomarker

Rheumatology key messages

- Machine learning algorithms provide accurate prediction of inadequate responders to MTX using simple clinical data on two training and one external validation cohort.
- We explored algorithm mechanisms of prediction that allow to identify a novel biomarker of inadequate response to MTX, namely high lymphocyte count.
- This study allows us to better predict with simple biomarkers patients with inadequate response to MTX who could benefit from other first-line therapeutic options.

Introduction

Despite the growing therapeutic arsenal in RA, MTX remains the first choice DMARD for RA patients in the international guidelines [1, 2]. In many cases, MTX the recommended first-line DMARD to use in a patient with early RA, effectively reduces disease activity, but still 30% to 40% of the patients respond inadequately, resulting in pain, irremediable joint destruction and disease progression. Accurate prediction of

Received: 26 August 2022. **Accepted:** 6 November 2022

inadequate response to MTX could ensure enable use of efficient second-line therapy such as MTX in combination therapy like triple conventional synthetic DMARD or targeted DMARDs [2].

Some factors have already been associated with MTX inadequate response (MTX-IR), such as female gender, current smoker, younger age or tender joint count [3–5]. Several studies tried identifying biological or genomic markers of MTX-IR, but none emerged as a reliable predictive factor so far [6].

With the increasing amount of available data, machine learning adoption in healthcare has been growing over the years [7]. Machine learning models learn patterns from data and assume these will reproduce in the future. The algorithms identify patterns and rules without being explicitly programmed to do so. This is of particular interest in medicine to identify previously unknown biomarkers or combinations of markers. Machine learning is now widely used in healthcare, especially in radiology and oncology, for diagnosis [8], prognosis or treatment recommendations [9].

In rheumatology, recent initiatives used genetic data and machine learning to predict response to MTX [10–12]. However, implementing these models in clinical practice remains a challenge because genetic data are unavailable in usual practice. Other approaches only used biological and clinical data but lack external validation in independent cohorts [13]. Finally, most of the studies implemented a black box approach that poorly describes how the patient characteristics contribute to the final model predictions and constrain the physician to resort to a medical device, limiting the clinical usability of such models [14].

In this study, rather than using machine learning as a black box device solving a clinical challenge, we used machine learning as a cutting-edge data analysis technique to unveil information hidden in the data. Understanding the key patient characteristics impacting the model's predictions provides precious insights to guide biomedical research.

This study builds a machine learning model to predict the therapeutic response to MTX after 9 months (± 6 months) in RA patients included in two independent cohorts – respectively, the ESPOIR and Leiden EAC – based on routinely available clinical and biological data. The objectives are 3-fold. First, we aimed to assess capabilities of machine learning to predict response to treatment from routine clinical data. Second, we evaluated such models' usefulness and potential impact in clinical practice. Third, we extracted a simple biomarker-based rule from the algorithm to easily identify inadequate responders. The performances of both the model and the simple rule were assessed for validation on data from a third external and independent clinical trial, namely the tREACH.

Patients and methods

Patients

Developing prognostic criteria and tools requires developing the model on a training dataset and validating the results on a validation set independent from the training base. The more different the validation set is, the better we can assess the model's generalization. For the training dataset, we included RA patients from two longitudinal and prospective early arthritis cohorts: ESPOIR, a French multicentric cohort [15], and the Leiden Early Arthritis Clinic cohort (Leiden EAC) [16]. ESPOIR and Leiden EAC are observational studies where RA patients are treated according to each centre's clinical practice. As external validation cohort, we used the tREACH trial [17], a randomized clinical trial comparing a triple DMARD therapy with MTX monotherapy in combination with low-dose glucocorticoid in newly diagnosed, DMARD naive, RA patients. Only RA patients with MTX monotherapy plus potential glucocorticoid bridging therapy were selected. The different nature of training and validation cohorts helped assess the models' ability to generalize on unseen and heterogeneous data. tREACH being a clinical trial, therapeutic response is assessed more precisely at constant time points; this helped assess algorithm performances more accurately, despite tREACH being the smallest data set. Moreover, as described below, we used cross-validation on the training set to prevent overfitting, which limits the need for a large validation set.

Patients were included if they fulfilled the 2010 ACR/EULAR criteria [18, 19] and received at least one dose of MTX (oral or SC in ESPOIR and LEAC, oral only in tREACH) and MTX monotherapy at any time of their disease; thus, combination therapies with other csDMARDs were excluded. We included a binary variable as models' input to account for corticosteroid treatment along with MTX. Patients stopping MTX within 6 months after initiation due to pregnancy, surgery, poor compliance to the protocol, or unknown reasons were excluded from the study population. In observational cohorts, if a patient received several distinct MTX therapeutic sequences, all of them were included in the dataset, provided they fulfilled the criteria listed above. On the other hand, in the validation dataset, each patient had a unique MTX therapeutic sequence.

End points

The first end point of our study was the prediction of the therapeutic response, defined as EULAR response [20] assessed 9 months (± 6 months) after treatment initiation. The EULAR response criteria classifies individual patients as non, moderate or good responders, dependent on the extent of change and the level of disease activity reached. In the study, we considered both good and moderate EULAR responders as responders. In the observational cohorts (ESPOIR and Leiden EAC), patients stopping MTX treatment within 6 months after initiation due to inefficacy or adverse events were considered inadequate responders (inadequate-responder imputation). Conversely, patients stopping MTX treatment within 6 months after initiation due to remission were considered responders. In the tREACH trial, a treat-to-target strategy was adopted. When the therapeutic target was not reached at 3 months, treatment was intensified, even if a good or moderate response was obtained according to EULAR criteria. If the patients included in tREACH had treatment change or intensification at 3 months despite achieving EULAR response, we considered them as responders to MTX. This end point is binary (inadequate response *vs* response) and is evaluated using the area under the curve ROC (AUROC). A higher AUROC corresponds to a better model.

To validate our biomarkers' results with other endpoints, we analysed if these identified biomarkers could also be predictors for reaching a state of low disease activity (LDA, DAS28 < 3.2) or DAS28-remission (DAS28 < 2.6).

Models and variables

The whole methodology process is illustrated in Supplementary Fig. S1, available at *Rheumatology* online.

The variables included in the model are demographic, clinical and biological measures available in routine clinical practice.

To deal with missing data, we compared multiple methods on the training set and selected the one that yielded the best results according to our evaluation criteria. The compared methods were mean imputation, median imputation, k-nearest-neighbors-based imputation (KNN) [21] and MICE (multiple imputation by chained equations) [22].

A variable selection process was applied to all available variables to select the most predictive features the models will use. An algorithmic variable selection process lets the models choose the most essential features independently of the physicians' input. Without supervision, machine learning algorithms can identify known response factors and potentially unveil new biological and clinical markers. To perform this process, we used the recursive feature elimination [23], described in Supplementary Data S1, available at *Rheumatology* online.

Four machine learning models were assessed: a logistic regression model, a random forest model [24], and two gradient boosted trees models. The Python library Scikit-Learn [25] was used to implement the regression and random forest models and the KNN and MICE imputation methods. LightGBM [26] and CatBoost [27] libraries were compared for the gradient boosted trees.

Each model uses the variables available at the last check-up before the beginning of the MTX therapeutic sequence to predict the outcome. The models output a probability of response which is compared with a decision threshold to finally obtain the binary therapeutic response. Patients with a probability of response above 0.5 were labelled as responders, while patients with a probability lower than 0.5 were labelled as inadequate responders.

Evaluation

The process (recursive feature elimination, data augmentation and imputation, and model training) was evaluated using a 5-fold cross-validation on the training dataset. The cross-validation process is detailed in Supplementary Data S1, available at *Rheumatology* online. Cross-validation is the first way to ensure our model doesn't overfit and allows for unbiased model and feature selection. Only the best trained model was evaluated on the external validation dataset to ensure the replication of the results. We computed formal external validation sample size using methods detailed in [28] based on cross-validated results. Calculations are detailed in Supplementary Data S1, available at *Rheumatology* online. We use the AUROC to compare the models and statistical comparisons are described in Supplementary Data S1, available at *Rheumatology* online.

Traditional epidemiological metrics exist when predicting therapeutic response. We computed, to provide clinical perspective: the positive predictive value (PPV), the negative predictive value (NPV), the sensitivity and the specificity.

We evaluated the model bias and applicability using the PROBAST framework; the complete form is available in Supplementary Data S2, available at *Rheumatology* online.

Explainability of the predictions

We studied how the models yielded their decisions to identify biomarkers of response to MTX. One of the most popular packages to date to explain machine learning predictions is SHAP (Shapley Additive exPlanations) [29]. SHAP is based on the concept of Shapley value which is specific to a patient and a characteristic. This value measures the weight of a patient's characteristic on the patient's predicted outcome. A positive (resp. negative) Shapley value indicates a positive (resp. negative) influence on patient response to treatment. Higher Shapley values indicate stronger influences on patient response and vice versa.

We performed Shapley explanation on the training set (ESPOIR and Leiden EAC) to identify profiles of inadequate responders based on the patient characteristics. We displayed explanation diagrams at the dataset level, plotting for each patient and each feature the contribution of this feature to the prediction. To find patients with similar prediction explanations, we performed clustering on Shapley values using the Kmeans algorithm [30] on the training set. An optimal number of clusters was obtained using the elbow method.

Given the small size of the external validation set, biomarker hypotheses were validated using odds ratio with a 90% confidence interval.

Ethics approval

The protocol of the ESPOIR Cohort study was approved in July 2002 by the ethical committee of Montpellier. The data is registered as ClinicalTrials.gov NCT03666091. For Leiden EAC, ethical approval was obtained from the 'Commissie Medische Ethiek' (medical ethics committee) of the Leiden University Medical Centre (B19.008). All participants provided written informed consent. The tREACH trial was approved by the medical ethics committees of the eight participating centres. It was registered as study ISRCTN26791028 by ISRCTN Registry. The respective scientific committees of the three cohorts approved the use of the data for this study.

Patient and public involvement

Patients were not involved in the design of this study. Patients will be informed of the results of this study by publication on the website and newsletter of the ESPOIR cohort.

Results

Screening process

In the training dataset, 674 patients with 732 therapeutic sequences were included; 493 therapeutic sequences came from the ESPOIR cohort and 239 from the Leiden EAC. A total of 674 sequences out of 732 (92%) were therefore MTX-naive. In the external validation dataset (tREACH), 138 patients were included. Conversely to other classical methodological approaches where 50/50 size is recommended between discovery and validation sets, machine learning recommends having a validation set size of around 20% of the total dataset when using cross-validation [31], which is what is achieved in this study. Total number of sequences above 3, 6 and 9 months is detailed in Supplementary Table S1, available at *Rheumatology* online, for each cohort.

The baseline characteristics of RA patients at the beginning of the treatment sequences are displayed in Table 1. Statistics on missing data are detailed in Supplementary Table S2, available at *Rheumatology* online. In the training dataset, 40% of treatment sequences were considered inadequate responses, while 30% of patients were responders in the external

Table 1. Characteristics of the training dataset (ESPOIR and Leiden EAC) and external validation dataset (tREACH) cohorts at the last visit before treatment initiation

	Training		Validation
Feature's name	**ESPOIR (*n* = 493)**	**EAC (*n* = 239)**	**tREACH (*n* = 138)**
Age, year	50 (12)	58 (14)	55 (14)
Female, *n* (%)	374 (76%)	143 (59%)	99 (72%)
Body mass index	25.2 (4.7)	25.5 (4.8)	26.3 (4.9)
DAS28	4.7 (1.6)	4.7 (1.2)	4.7 (1.3)
CRP, mg/L	20 (33)	20 (24)	19 (26)
ESR, mm	28 (24)	32 (24)	27 (22)
Creatininemia, μmol/L	73 (14)	71 (16)	72 (17)
AST, UI/L	22 (9)	NA	23 (13)
ALT, UI/L	23 (14)	NA	26 (21)
White blood, cells/mm^3	7.6×10^3 (2.4×10^3)	8.5×10^3 (2.2×10^3)	8.5×10^3 (2.6×10^3)
Neutrophils, cells/mm^3	4.9×10^3 (2.1×10^3)	5.9×10^3 (2.1×10^3)	5.5×10^3 (2.1×10^3)
Lymphocytes, cells/mm^3	1.8×10^3 (7.3×10^2)	1.8×10^3 (6.3×10^2)	2.1×10^3 (7.7×10^2)
ACPA, *n* (%)	274 (55%)	148 (61%)	72 (50%)
Rheumatoid factor, *n* (%)	294 (60%)	162 (67%)	67 (47%)
Tender joints count	7.8 (7.3)	7.0 (5.7)	6.8 (5.5)
Swollen joints count	6.1 (5.6)	5.7 (4.9)	6.7 (5.3)
corticosteroid co-treatment, *n* (%)	191 (37%)	128 (49%)	86 (62%)
Global health evaluation	54 (28)	43 (25)	51 (23)
Ever smoked, *n* (%)	238 (48%)	167 (72%)	96 (70%)
Current smoking, *n* (%)	100 (20%)	64 (26%)	35 (28%)
Cumulative smoking dose, pack years	9.0 (14)	21 (15)	15 (18)
HAQ	0.94 (0.70)	1.0 (0.67)	1.09 (0.67)
Responders, *n* (%)	277 (56%)	161 (66%)	96 (70%)

Results are presented as follows: mean (s.d.) for continuous variables and amount (percentage) for binary variables. NA, not available.

validation set. Leiden EAC does not contain AST, ALT. These variables were imputed with the missing data imputation methods presented in the methods.

Variable selection and models evaluation

From the 22 variables included, eight were selected by the model for their highly predictive value to predict the EULAR response. They were: DAS28, creatininemia, leucocytes, lymphocytes, AST, ALT, swollen joints count and corticosteroids co-treatment.

The four machine learning models were assessed for the prediction of the therapeutic response. The performance of each model and their comparisons are displayed in Supplementary Fig. S2 and Supplementary Table S3, available at *Rheumatology* online. LightGBM with the MICE missing value imputer performed the best on the training dataset with an AUROC of 0.73 [95% CI (0.62, 0.74)]. The results replicated well on the external validation set with an AUROC of 0.72 [95% CI (0.63, 0.80)]. ROC curves are presented in Fig. 1, and calibration curves are presented in Supplementary Fig. S3, available at *Rheumatology* online.

We computed the number of predicted responders and the accuracy, sensitivity, specificity, PPV and NPV on the external validation set (Table 2). In the external validation cohort, 70% of patients were responders to MTX. The model predicts response with an accuracy of 74% (95% CI 67%, 81%). Patients predicted as inadequate responders had 38% [95% CI (20%, 58%)] chance to respond. Using the algorithm to decide to initiate MTX would decrease inadequate-response rate from 30% to 23% [95% CI (17%, 29%)]. The model identifies 38% (=Specificity, 95% CI: 24%, 53%) of the inadequate responders.

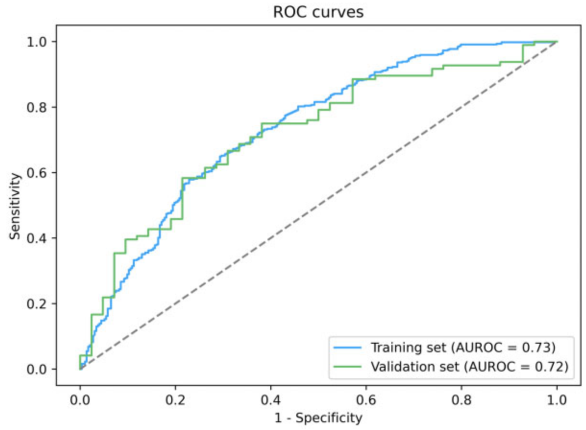

Figure 1. ROC curves for the best model (LightGBM) on the training set (ESPOIR + Leiden EAC) and the external validation set (tREACH)

Subgroups of inadequate responders with moderate or high disease activity

We next investigated how each of the eight variables selected by the model (DAS28, creatininemia, leucocytes, lymphocytes, AST, ALT, swollen joints count and corticosteroids co-treatment) impacted the prediction. Shapley values measure for a given sequence the influence of a given variable on response. When displaying Shapley values, DAS28 is the variable that has the strongest influence on the response (Fig. 2). As expected, patients with a low DAS28 (inferior to 3.2 on Fig. 2) are unlikely to reach EULAR response and have highly negative Shapley values associated with these DAS28 values. Those patients with a DAS28 inferior to 3.2 are the least interesting because they are already in a low disease activity state, so we focused on patients with moderate or high disease activity (DAS28 > 3.2).

Table 2. AUROC (95% CI), number of predicted responder sequences (% of the total dataset), and metrics (95% CI)

Training set (732 sequences)	Validation set (138 patients and sequences)						
AUROC	AUROC	Patients predicted as responders	Accuracy	Sensitivity	Specificity	PPV	NPV
73% (64%–75%)	72% (63%–80%)	112 (81%)	74% (67%–81%)	90% (83%–95%)	38% (24%–53%)	77% (69%–84%)	62% (42%–80%)

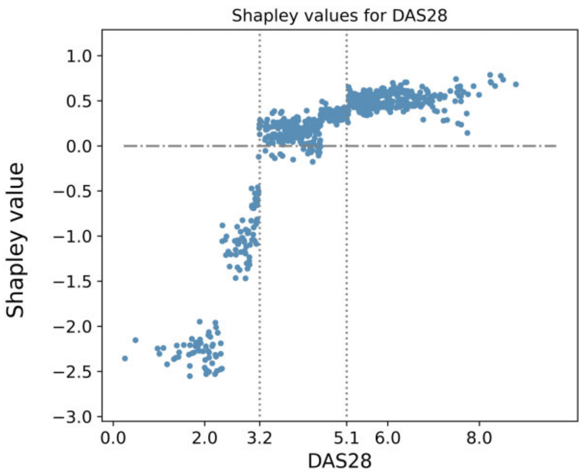

Figure 2. Shapley values for DAS28 computed on the training set. Each dot represents a sequence; a positive Shapley value means the feature contributed towards response, and the higher the value, the more it contributed

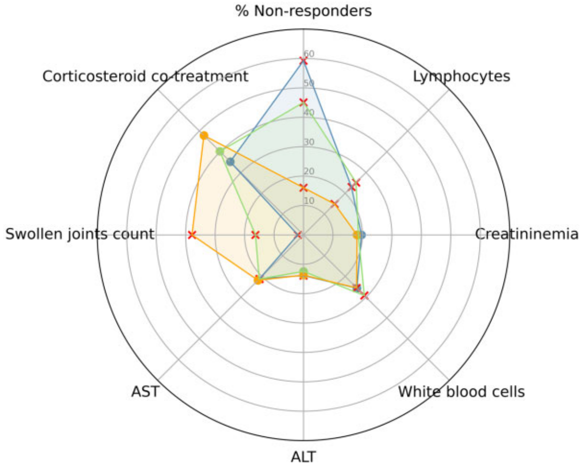

Figure 3. Radar plot showing the mean biomarker values for each cluster on the training set (blue cluster: low response; green cluster: standard response; orange cluster: good response). The axes show values expressed as percentages with respect to the maximum value for each biomarker. When the mean of a characteristic in a cluster is significantly different from the sequences outside of the cluster, the value is marked with a red cross instead of a dot

Using the K-means algorithm to cluster Shapley values and find groups of patients whose predictions were explained similarly, we obtained three distinct clusters of patients with moderate or high activity (Supplementary Fig. S4, available at *Rheumatology* online). Displaying the characteristics of each group (Fig. 3) shows a high response rate in cluster 3 (in orange) (80%). In comparison, cluster 2 (in green) is average (45% of inadequate responders), and cluster 1 in blue has a high rate of inadequate responders (60%). The patients in cluster 3 (orange) had average values for most characteristics except for a much lower blood lymphocyte count and a higher swollen joint count (SJC).

Thus, two items are good candidates to be biomarkers of inadequate response to MTX: the lymphocyte and swollen joints counts. Patients with high lymphocyte count and low swollen joint count seem to be less likely to respond (blue cluster). We computed two thresholds on the training set (ESPOIR and Leiden EAC) to identify simple rules of inadequate response regarding these two markers and validated these rules on the external validation set. RA patients with a lymphocyte count higher than 2000 cells/mm^3 or with a number of swollen joints of 0 or 1 are much less likely to reach EULAR response (Table 3).

LDA and DAS28-remission endpoints for high lymphocytes subgroup

To extend subgroup analysis beyond EULAR response, we analysed if high lymphocyte count was a predictor of reaching a low disease activity (LDA: DAS28 $<$ 3.2) or DAS28-remission state (DAS28 $<$ 2.6). The threshold for these endpoints was computed on the training set, and results suggest that patients with a lymphocyte count superior to 2000 cells/mm^3 were less likely to reach these endpoints (Supplementary Table S4, available at *Rheumatology* online). On the external validation set, 21 patients out of 55 (38%) reached an LDA state in the high lymphocyte count subgroup, compared with 58 out of 123 in the moderate to high activity population, giving an odds-ratio of 0.52 [90% CI (0.28; 0.95)]. Regarding DAS28-remission, 12 out of 55 (22%) reached a DAS28-remission state in the high lymphocyte count subgroup, compared with 35 out of 123 in the moderate-to-high activity population, giving an odds ratio of 0.55 [90% CI (0.28–1.08)] which suggests the same effect but lacks statistical power to be conclusive.

We did not include swollen joints count in this analysis as a low swollen joints count correlates with a lower initial DAS28 and thus higher likelihood of reaching LDA or DAS28-remission.

Discussion

This study established that machine learning models efficiently assess the therapeutic response to MTX using exclusively data available in clinical routine. We obtained a good AUROC on both the training and the external validation cohort. Analysing how each variable impacted the model predictions, we identified a novel simple rule (lymphocytes >2000/mm^3) that can be used without the machine learning algorithm to identify MTX inadequate responders in the

Table 3. Metrics of interest for each subgroup computed on the validation set on patients with moderate and high activity

Subgroups	Number of patients in the group	Number of responders in the group (%)	EULAR response odds ratio (90% CI)
All patients with moderate or high activity (DAS > 3.2)	123	88 (72%)	Odds reference: 2.57
High lymphocyte count subgroup (>2000 cells/mm^3)	55	35 (64%)	0.5 (0.25–0.96)
Low swollen joints count subgroup (0 or 1)	14	6 (42%)	0.25 (0.09–0.64)
High lymphocyte count subgroup or low swollen joint count	61	46 (59%)	0.28 (0.14–0.56)

population of RA patients with a DAS28 above 3.2 at baseline. Interestingly, this population with moderate or high clinical activity is the most common population in which MTX is initiated after an RA diagnosis. A recent study that combined genetic and clinical data outlined the difficulty to predict response, reaching poor replication in the external validation cohort due to this population of mid-range DAS28 values [10].

The study's first objective was to assess the performances of a machine learning model based exclusively on clinical and biological data available in routine clinical practice. It is worth noting that the results replicate properly between the training set (ESPOIR + Leiden EAC) and the external validation set (tREACH), with an AUC of 0.72 compared with 0.73 in the training dataset. This is remarkable because ESPOIR and Leiden EAC are observational studies while tREACH is a clinical trial. This is to be compared with AUROC in the range of 0.6–0.7 using genomic data in the DREAM RA challenge [32, 33], or studies using only biological data to predict response to TNF inhibitors [34]. Performances obtained by machine learning models were slightly better than conventional methods such as logistic regression, which is consistent with previous findings [35].

The study's second objective was to assess the potential of machine learning-derived tools in clinical practice. In the external validation set, our model could enable clinicians to skip the MTX first line for predicted non-responders and switch to another treatment with around 62% confidence in inadequate response to MTX (NPV), compared with the 30% of inadequate responders. This would decrease the inadequate-response rate from 30% to 23%. Reaching this confidence level using our algorithm would result in 19% fewer patients treated with MTX because they are predicted not to respond. Given the price difference that currently persists between MTX and most other treatments and the augmented risk of side effects, this is probably not enough for introducing a tDMARD as first line but could be enough to start a triple csDMARD therapy. It is a first step towards precision medicine, and this 62% NPV can be improved by future studies. Using our model in clinical practice would allow physicians to prescribe MTX with a confidence of 77% (PPV) compared with the 70% of responders in the tREACH trial. It is a significant improvement compared with current clinical practice. Those results need to be validated in a broader study. The validation set size is indeed sufficient regarding the observed/expected ratio computed in Supplementary Data S1, available at *Rheumatology* online (135 patients minimum) but insufficient to estimate accurate calibration slope (477 patients minimum), very important to estimate usefulness of a predicting algorithm in clinical practice.

The third objective of the study was to unveil inadequate-responding patients' subgroups by exploring the algorithm's inner decision mechanisms. Contrarily to linear methods, machine learning algorithms learn complex relationships between variables, which can yield new medical insights. This study thoroughly details the explainability of the algorithm using Shapley values. Our method was inspired by the successful application of Shapley values in oncology [36] and recently in rheumatology [37, 38]. Logically, DAS28 plays a key role in predicting the EULAR response, as the EULAR criteria is biased and favours the response of high-DAS28 patients. Our model identified two subgroups of inadequate responders: patients with moderate or high disease activity and high lymphocyte count or a small number of swollen joints. While it is already well known that patients with few swollen joints but several tender joints tend to have a worse therapeutic response [39], high blood lymphocyte count appears as a new reliable biomarker to predict MTX-IR. Patients with high or moderate disease activity and a lymphocyte count above 2000 cells/mm^3 have are significantly more likely of being inadequate responders, using the EULAR response. We confirmed these results using other endpoints. High lymphocyte count predicts less chances of reaching LDA but DAS28-remission failed to reach statistical significance.

Only hypothesis can be made at this stage for linking high lymphocytes level and MTX-IR. The most commonly accepted mechanism of action of MTX is a role in increasing the levels of adenosine, which has immunomodulatory properties [40, 41]. Thus, higher lymphocyte counts could be the hallmark of higher activation and thus be harder to control by MTX-induced production. It would be interesting to design a study of prediction of response to MTX with an evaluation of subsets of lymphocytes and to assess if this biomarker is also predictive of inadequate response to other treatments.

Our results may have important clinical consequences as patients in moderate or high activity (DAS28 > 3.2) are today the most numerous patients in whom MTX is started, and blood lymphocyte count is part of the routine blood measurements in these patients. Even if high blood lymphocyte count appears predictive of worse response to MTX in two cohorts and one clinical trial, the validity of this new simple biomarker must be confirmed in a broader study with more patients.

This study faces several limitations. As the study's goal was to predict treatment effect, we chose the EULAR response as the primary end point, despite clinicians more often looking at a state of low disease activity or remission in practice. However, blood lymphocyte count was also predictive of a lower rate of patients achieving LDA. Another limitation is the variability in clinical practice and measurement time points of the ESPOIR and Leiden EAC data inherent to observational studies. This variability results in a large window of evaluation (9 months ± 6 months after MTX initiation) to include more patients in the training dataset, as the timespan between treatment initiation and the next visit can be long in those observational cohorts. However tREACH being a clinical trial, performances were assessed much more accurately on the external validation set where the response was systematically evaluated at 3 and 6 months. Moreover, the training dataset and external validation are of different natures as ESPOIR and LEAC are observational cohorts while tREACH is a clinical trial, with a predefined therapeutic strategy including a higher use of glucocorticoids. We see this difference as a strength of our study as results replicate well and highlight the model's ability to generalize. Despite different patient populations, we chose to impute missing variables to be more comprehensive in potential predictors, which could lead to introducing bias for variables largely missing such as AST and ALT in LEAC. Another study should confirm the findings of the present regarding predictors of response to MTX. To include more data in the training set, several therapeutic sequences of the same patient could be included, which results in a lack of independence. This was accounted in cross-validation, and we prevented having sequences from the same patient in the training set and validation holdout, but this can still yield a small bias in training. Ultimately, data on ethnicities is not directly available in these cohorts, but most of the studied population was Caucasian. This introduces bias in our model, which should be measured in future work on different cohorts.

In conclusion, our study highlights the usefulness of machine learning in unveiling simple biomarkers such as high blood lymphocyte count to identify subgroups of inadequate responders to MTX to guide new therapeutic strategies. Further work is needed to validate this approach and to test this simple biomarker in treatment strategy trials to improve the outcome of MTX-IR patients.

Supplementary data

Supplementary data are available at *Rheumatology* online.

Data availability statement

Data used in this study are available upon agreement of the scientific committees of each of the cohorts, namely ESPOIR, Leiden EAC and tREACH trial.

Contribution statement

V.B., J.D., S.B. and X.M. conceptualized the study and performed the data request. V.B. and J.D. built the prediction models, analysed the data, and prepared all figures. S.B and X.M. assisted with the interpretation of the data. P.H.C. provided expert guidance for all aspects of the study. V.B., J.D., S.B. and X.M. wrote the manuscript. All authors reviewed the final manuscript.

Funding

P.H.C., B.F., F.G., P.H.P.dJ., J.W.H., M.V., A.H.M.vdH.M. and X.M. have no financial disclosures. V.B. and J.D. are Scienta Lab employees and hold Scienta Lab stocks. S.B. was supported by FHU CARE.

Disclosure statement: The authors have declared no conflicts of interest.

Acknowledgements

For the first 5 years of the ESPOIR cohort, an unrestricted grant from Merck Sharp and Dohme (MSD) was allocated. Two additional grants from INSERM were obtained to support part of the biological database. The French Society of Rheumatology, Pfizer, AbbVie, Lilly, and more recently Fresenius and Biogen supported the ESPOIR cohort study. We also thank Nathalie Rincheval (Montpellier), who did expert monitoring and data management, and all the investigators who recruited and followed the patients. We also thank investigators from Leiden EAC and tREACH studies for collecting and providing the data for this study.

References

1. Fraenkel L, Bathon JM, England BR *et al.* 2021 American college of rheumatology guideline for the treatment of rheumatoid arthritis. Arthritis Care Res 2021;73:924–39.
2. Smolen JS, Landewé RBM, Bijlsma JWJ *et al.* EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. Ann Rheum Dis 2020;79:685–99.
3. Romão VC, Canhão H, Fonseca JE. Old drugs, old problems: where do we stand in prediction of rheumatoid arthritis responsiveness to methotrexate and other synthetic DMARDs? BMC Med 2013;11:17.
4. Saevarsdottir S, Wallin H, Seddighzadeh M *et al.* Predictors of response to methotrexate in early DMARD naïve rheumatoid arthritis: results from the initial open-label phase of the SWEFOT trial. Ann Rheum Dis 2011;70:469–75.
5. de Rotte MCFJ, Pluijm SMF, de Jong PHP *et al.* Development and validation of a prognostic multivariable model to predict insufficient clinical response to methotrexate in rheumatoid arthritis. Abu-Shakra M, ed. PLoS ONE 2018;13:e0208534.
6. Ling S, Bluett J, Barton A. Prediction of response to methotrexate in rheumatoid arthritis. Expert Rev Clin Immunol 2018;14:419–29.
7. Shah P, Kendall F, Khozin S *et al.* Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med 2019;2:69.
8. Aggarwal R, Sounderajah V, Martin G *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021;4:65.
9. Sammut S-J, Crispin-Ortuzar M, Chin S-F *et al.* Multi-omic machine learning predictor of breast cancer therapy response. Nature 2022;601:623–9.
10. Myasoedova E, Athreya AP, Crowson CS *et al.* Towards individualized prediction of response to methotrexate in early rheumatoid arthritis: a pharmacogenomics-driven machine learning approach. Arthritis Care Res 2022;74:879–88.
11. Gosselt HR, Verhoeven MMA, Bulatović-Calasan M *et al.* Complex machine-learning algorithms and multivariable logistic regression on par in the prediction of insufficient clinical response to methotrexate in rheumatoid arthritis. J Pers Med 2021;11:44.
12. Duong SQ, Crowson CS, Athreya A *et al.* Clinical predictors of response to methotrexate in patients with rheumatoid arthritis: a machine learning approach using clinical trial data. Arthritis Res Ther 2022;24:162.
13. Sergeant JC, Hyrich KL, Anderson J *et al.* Prediction of primary non-response to methotrexate therapy using demographic, clinical and psychosocial variables: results from the UK Rheumatoid Arthritis Medication Study (RAMS). Arthritis Res Ther 2018;20:147.
14. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17:195.

15. Combe B, Benessiano J, Berenbaum F *et al.* The ESPOIR cohort: a ten-year follow-up of early arthritis in France. Joint Bone Spine 2007;74:440–5.
16. de RD, van der LM, Knevel R, Huizinga TWJ, van der H-V. M A. Predicting arthritis outcomes–what can be learned from the Leiden Early Arthritis Clinic? Rheumatology 2011;50:93–100.
17. de JP, Hazes JM, Han HK *et al.* Randomised comparison of initial triple DMARD therapy with methotrexate monotherapy in combination with low-dose glucocorticoid bridging therapy; 1-year data of the tREACH trial. Ann Rheum Dis 2014;73:1331–9.
18. Aletaha D, Neogi T, Silman AJ *et al.* 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569–81.
19. Aletaha D, Neogi T, Silman AJ *et al.* 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann Rheum Dis 2010;69:1580–8.
20. Fransen J, van RP. The Disease Activity Score and the EULAR response criteria. Clin Exp Rheumatol 2005;23:S93–99.
21. Troyanskaya O, Cantor M, Sherlock G *et al.* Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17:520–5.
22. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J R Stat Soc 1960;22:302–6.
23. Zeng X, Chen Y-W, Tao C. Feature selection using recursive feature elimination for handwritten digit recognition. In: 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Kyoto: IEEE; 2009:1205–8. https://ieeexplore.ieee.org/document/5337549/ (29 July 2022, date last accessed).
24. Breiman L. Random forests. Mach Learn 2001;45:5–32.
25. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM; 2016:785–94. https://dl.acm.org/doi/10.1145/2939672.2939785 (15 October 2021, date last accessed).
27. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. *ArXiv181011363 Cs Stat*. 2018. http://arxiv.org/abs/1810.11363 (15 October 2021, date last accessed). preprint: not peer reviewed.
28. Riley RD, Debray TPA, Collins GS *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med 2021;40:4230–51.
29. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S *et al.*, eds. Advances in Neural Information Processing Systems. Vol. 30. Red Hook, NY: Curran Associates, Inc, 2017. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (15 October 2021, date last accessed).
30. Mannor S, Jin X, Han J *et al.* K-means clustering. In: C Sammut, GI Webb, eds. Encyclopedia of machine learning. Boston, MA: Springer US, 2011: 563–4. http://link.springer.com/10.1007/978-0-387-30164-8_425 (27 March 2022, date last accessed).
31. Hastie T, Tibshirani R, Friedman JH. Model assessment and selection. In: HastieT, Tibshirani R, Friedman JH, eds. Elements of statistical learning. New York City: Springer, 2001:222.
32. Sieberts SK, Zhu F, García-García J *et al.* Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. Nat Commun 2016;7:12460.
33. Guan Y, Zhang H, Quang D *et al.* Machine learning to predict anti–tumor necrosis factor drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers. Arthritis Rheumatol 2019;71:1987–96.
34. Bouget V, Duquesne J, Hassler S *et al.* Machine learning predicts response to TNF inhibitors in rheumatoid arthritis: results on the ESPOIR and ABIRISK cohorts. RMD Open 2022;8:e002442.
35. Westerlind H, Maciejewski M, Frisell T *et al.* What is the persistence to methotrexate in rheumatoid arthritis, and does machine learning outperform hypothesis-based approaches to its prediction? ACR Open Rheumatol 2021;3:457–63.
36. Moncada-Torres A, van MM, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep 2021;11:6968.
37. Queiro R, Seoane-Mato D, Laiz A *et al.* Minimal disease activity (MDA) in patients with recent-onset psoriatic arthritis: predictive model based on machine learning. Arthritis Res Ther 2022;24:153.
38. Angelini F, Widera P, Mobasheri A *et al.* Osteoarthritis endotype discovery via clustering of biochemical marker data. Ann Rheum Dis 2022;81:666–75.
39. Michelsen B, Kristianslund EK, Hammer HB *et al.* Discordance between tender and swollen joint count as well as patient's and evaluator's global assessment may reduce likelihood of remission in patients with rheumatoid arthritis and psoriatic arthritis: data from the prospective multicentre NOR-DMARD study. Ann Rheum Dis 2017;76:708–11.
40. Brown PM, Pratt AG, Isaacs JD. Mechanism of action of methotrexate in rheumatoid arthritis, and the search for biomarkers. Nat Rev Rheumatol 2016;12:731–42.
41. Bitoun S, Nocturne G, Ly B *et al.* Methotrexate and BAFF interaction prevents immunization against TNF inhibitors. Ann Rheum Dis 2018;77:1463–70.