Arthritis Research & Therapy

RESEARCH

Open Access

Predicting rheumatoid arthritis progression from seronegative undifferentiated arthritis using machine learning: a deep learning model trained on the KURAMA cohort and externally validated with the ANSWER cohort

Takayuki Fujii[1,2*], Koichi Murata[1,2], Hirohiko Kohjitani[3], Akira Onishi[1], Kosaku Murakami[4], Masao Tanaka[1], Wataru Yamamoto[5], Koji Nagai[6], Ayaka Yoshikawa[6], Yuki Etani[7], Yasutaka Okita[8], Naofumi Yoshida[9], Hideki Amuro[9], Tadashi Okano[10], Yo Ueda[11], Takaichi Okano[12], Ryota Hara[13], Motomu Hashimoto[14], Akio Morinobu[15] and Shuichi Matsuda[2]

Abstract

Background Undifferentiated arthritis (UA) often develops into rheumatoid arthritis (RA), but predicting disease progression from seronegative UA remains challenging because seronegative RA often does not meet the classification criteria. This study aims to build a machine learning (ML) model to predict the progression from seronegative UA to RA using clinical and laboratory parameters.

Methods KURAMA cohort (training dataset) and ANSWER cohort (validation dataset) were utilized. Patients with seronegative UA were selected based on specific inclusion and exclusion criteria. Clinical and laboratory parameters, including demographic data, acute phase reactants, autoantibodies, and physical examination findings, were collected. Various ML models, including a Feedforward Neural Network (FNN), were developed and compared. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, and other metrics. SHapley Additive exPlanations (SHAP) values were computed to interpret the importance of variables.

Results KURAMA cohort included 210 patients with seronegative UA, of whom 57 (27.1%) progressed to RA. The FNN model demonstrated the highest predictive performance with an AUC of 0.924 and a sensitivity of 80.7% in the training dataset. Validation with ANSWER cohort (140 patients; 32.1% progressed to RA) showed an AUC of 0.777, sensitivity of 77.8%. MMP-3 had the highest impact on the model.

*Correspondence:
Takayuki Fujii
fujiit@kuhp.kyoto-u.ac.jp

Full list of author information is available at the end of the article

Conclusions The FNN model exhibited robust performance in predicting the progression of RA from seronegative UA and maintained substantial sensitivity in an independent validation cohort. This model using only clinical and laboratory parameters has potential for predicting RA progression in patients with seronegative UA.

Introduction

Rheumatoid arthritis (RA) poses a significant burden on patients, presenting a wide range of clinical manifestations, including joint arthritis and extra-articular symptoms, which can lead to substantial morbidity and disability if left untreated or inadequately managed [1, 2]. Early diagnosis and treatment are essential for RA, as delayed diagnosis worsens treatment response and joint prognosis [3, 4]. The American College of Rheumatology (ACR)/European League Against Rheumatism (EULAR) 2010 RA classification criteria have been proposed to enable RA diagnosis in its early stages [5]. However, patients with recent onset inflammatory arthritis often do not always meet ACR/EULAR 2010 RA classification criteria and include a broad range of disease entities [6].

Within the heterogeneous group of early-stage inflammatory joint disorders, undifferentiated arthritis (UA) refers to inflammatory arthritis that does not fulfill the diagnostic criteria for specific arthropathies such as RA [6, 7]. This diagnostic uncertainty presents a clinical challenge, as predicting the disease course and optimal management strategies for UA remain difficult. Furthermore, UA has been associated with an increased risk of progression to defined rheumatic diseases, making accurate identification and timely intervention crucial [8–11]. A model predicting the evolution from UA to RA using clinical and laboratory parameters has shown good performance in estimating the risk of developing RA in patients with UA [12, 13]. However, after the introduction of the ACR/EULAR 2010 RA classification criteria, the characteristics of UA patients have changed [11]. Therefore, a prediction tool for estimating the clinical course of UA patients who do not meet RA criteria is beneficial for both patients and clinicians.

Rheumatoid factor (RF) and anti-citrullinated protein antibody (ACPA) are manifestations of autoimmunity in RA [14]. However, seronegative RA, which is negative for RF and ACPA, accounts for approximately 30% of RA and is often misdiagnosed because few patients meet ACR/EULAR 2010 RA classification criteria [15, 16]. In addition, RF and ACPA are risk factors for developing RA, and therefore individuals who are positive for RF and ACPA are considered to be at "high risk" [17–19]. This highlights the importance of predicting the progression to RA in seronegative UA patients, in whom RF and ACPA were negative at the first clinical evaluations. However, this prediction is difficult because approximately 50% of UA cases are self-limiting and about 30% evolve to RA [20–22].

Machine learning (ML), a subset of artificial intelligence, makes use of algorithms and statistical models to interpret and analyze complex data sets [23–25]. ML has demonstrated significant potential in various domains, including disease prediction, patient stratification, and personalized medicine [23–25].

In this study, we aimed to build a machine learning (ML) model designed to predict the progression from UA to RA in seronegative patients. Our model utilized only clinical and laboratory parameters obtained in routine clinical settings. We also validated the model using an external dataset to ensure generalizability. Although uncertainty remains regarding the diagnosis of RA progression because the diagnostic criteria could not be standardized across hospitals and physicians in the cohorts due to the lack of consensus in diagnosing seronegative RA in its early stages [26, 27], we developed an externally validated ML model with good predictive performance for RA development from UA using real-world data, which may be valuable in daily rheumatology practice.

Methods

Patients

We used data from the KURAMA cohort as a training dataset and data from the ANSWER cohort as a validation dataset. The KURAMA cohort, established in 2011, is a single-center, observational cohort study of RA [28–30]. The ANSWER cohort, established in 2018, is a multicenter, longitudinal cohort study of RA involving 9 hospitals including Kyoto University [31–34].

Clinical inflammatory arthritis that did not fall into a specific diagnosis after initial clinical evaluation was classified as UA. In KURAMA cohort, we analyzed patients who were initially diagnosed as seronegative UA from 2011 to 2022. Exclusion criteria were as follows: (1) patients under 18 years of age, (2) patients with a history of malignancy, (3) patients with known autoimmune disease or treated with immunosuppressants, (4) patients with a history of RA diagnosis at another clinic, (5) patients not followed for more than 6 months even if UA persisted, (6) patients who were positive for RF and/or ACPA, and (7) patients with ACR/EULAR 2010 RA classification criteria ≥ 6 at the baseline. All UA patients were allowed to revisit our clinic even after the regular follow-up was suspended.

In ANSWER cohort, patients who were initially diagnosed as seronegative UA were analyzed after excluding data that met the following conditions: (1) data from the

KURAMA cohort (the ANSWER cohort is a multicenter cohort that includes the KURAMA cohort), (2) patients with ACR/EULAR 2010 RA classification criteria ≥ 6 at the baseline, (3) patients treated with immunosuppressants or with known autoimmune disease.

Definition of progression to RA

The progression to RA was not always determined based on ACR/EULAR 2010 classification criteria [27]. In the KURAMA cohort dataset, patients' medical records were retrospectively analyzed and the reason for diagnosis was collected. When patients were diagnosed with diseases other than RA, they were not included in RA progression group even if the patients fulfilled ACR/EUALR 2010 classification. In the ANSWER cohort dataset, RA progression was identified from the database.

Clinical and laboratory parameters

At the baseline, we obtained patients' demographic and anthropometric data, smoking history, family history of RA. We also collected baseline acute phase reactants (C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR), autoantibodies (RF and ACPA), and matrix metalloproteinase 3 (MMP-3). Physical examinations were performed, and 28 tender joint counts (TJC), 28 swollen joint counts (SJC), physician global assessment (PhGA), and patient global assessment (PtGA) were obtained. Based on these data, 2010 ACR/EULAR RA classification criteria points [5] and the clinical disease activity index (CDAI) [35] were evaluated. Functional disability was also assessed using the Health Assessment Questionnaire Disability Index (HAQ-DI) [36].

Statistical analyses and machine learning model building.

Wilcoxon rank sum test, chi-square test with Yates's correction, Receiver Operating Characteristic (ROC) curve analysis, and ML modeling were performed using Scipy v1.11.4, Scikit-learn v1.4.1 in Python v3.8.16. Kaplan-Meier analyses were performed using lifelines v.0.28.0. PyCaret v2.3.10 was used to compare ML models except for Feedforward Neural Network (FNN). FNN was built using Tensorflow v2.8.0 and Keras v.2.8.0. SHapley Additive exPlanations (SHAP) values were computed using shap v0.46.0. The highest proportion of missing values was observed in MMP-3 and CDAI (3.3%) in the KURAMA cohort and in MMP-3 and HAQ-DI (7.14%) in the ANSWER cohort, and summarized in Supplementary Table 1. We imputed missing values using multiple imputations by chained equations (MICE) and generated 100 imputed datasets using random forest in R v4.3.3.

Models predicting progression from UA to RA were trained using the KURAMA cohort and externally validated using the ANSWER cohort. Performance measures included area under the receiver operating characteristic curve (AUC), positive prediction value (PPV), sensitivity, specificity, accuracy, and F1 score, which is a measure of the harmonic mean of sensitivity and PPV and were computed using 5-fold cross validation (Fig. 1A). Hyperparameter tuning for ML models except FNN was performed using "tune_model" function of PyCaret. Learning rate was tuned for FNN (Supplementary Fig. 1A). Sample size evaluation was performed by drawing learning curves with different number of samples (Supplementary Fig. 1B) [37]. Values were not normalized in the dataset because better performance was obtained without normalization for both FNN and other ML models. The p-value threshold for statistical significance was set at < 0.05.

Study approval and design

The institutional review board at all hospitals participating in the KURAMA and the ANSWER cohorts approved this study. Written informed consent was obtained from all participants according to the Declaration of Helsinki. We reported the study according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline [38].

Results

Patient characteristics

We identified 519 patients with initial diagnosis of UA (including seropositive patients) in the KURAMA training cohort and analyzed 210 seronegative UA patients after applying exclusion criteria (Supplementary Fig. 2). Among the 210 patients with seronegative UA in the KURAMA cohort, 57 patients (27.1%) progressed to RA (Table 1). Sonography and MRI were performed in 56 (26.7%) and 59 (28.1%), respectively. In 57 patients with RA progression, RA diagnosis was made based on presence of joint synovitis, tenosynovitis, bone edema, or bone erosion identified using sonography or MRI (46 patients), meeting ACR/EULAR criteria (6 patients), radiographic progression (3 patients), and clinical judgement (4 patients). Among 153 patients who did not develop RA, 105 patients remained UA or spontaneously resolved, 15 patients were diagnosed with osteoarthritis, 33 patients were diagnosed as other inflammatory arthropathies such as systemic lupus erythematosus, systemic sclerosis, or sarcoidosis (Supplementary Table 2). The median time to RA progression was 37 days in the RA group and the median time to diagnosis was 105 days in the non-RA group (Table 1).

We first compared clinical and laboratory parameters. Older age and higher CRP, ESR, MMP-3, PhGA, PtGA, TJC, SJC, and HAQ-DI were observed in patients who progressed to RA compared to those who did not develop

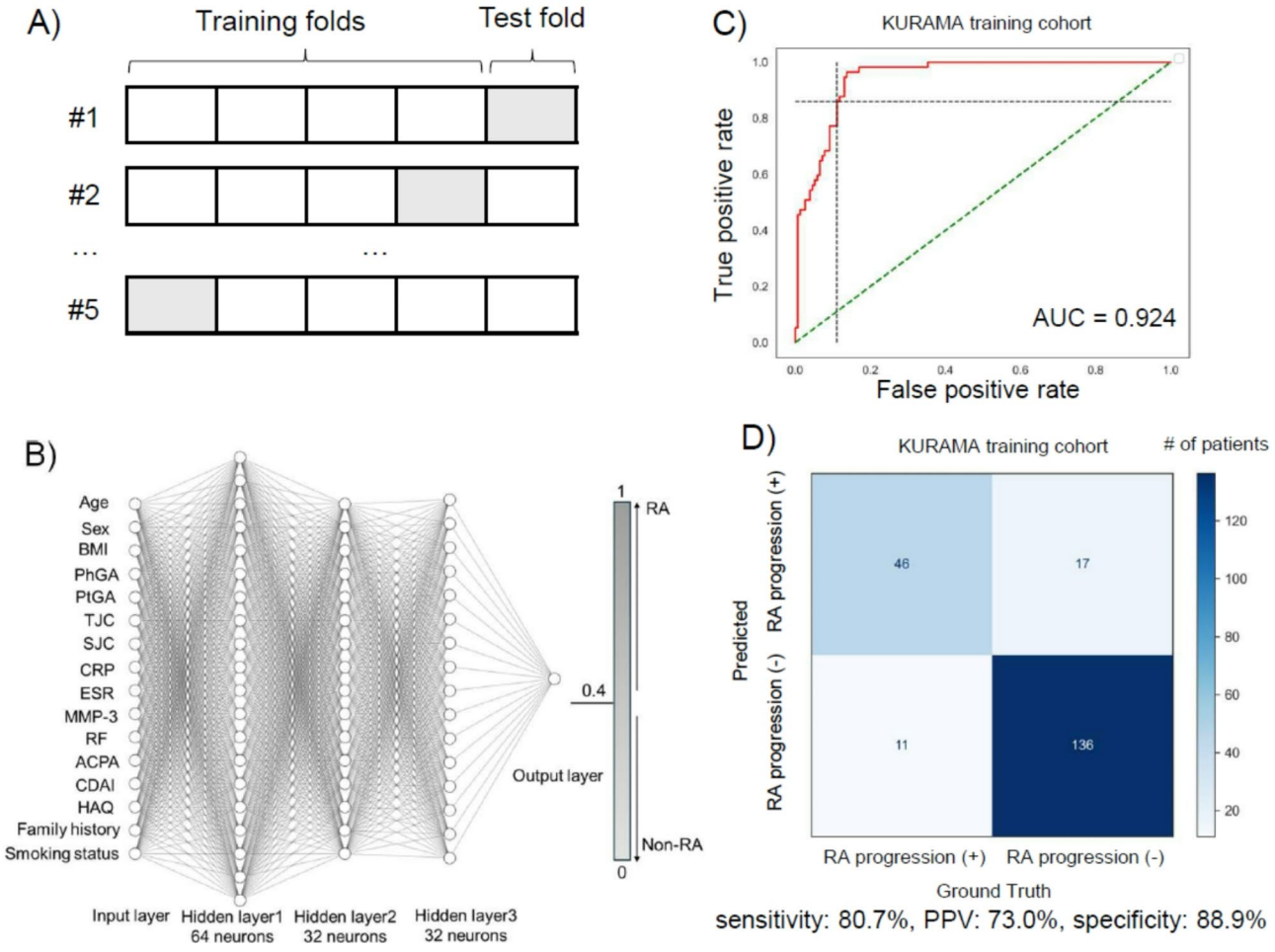

Fig. 1 Architecture of Feedforward Neural Network (FNN) and its performance. (**A**) Schematic explanation of 5-fold cross-validation. (**B**) Structure of FNN. An optimized threshold was set at 0.4 to predict RA progression. (**C**) Receiver operating characteristic curve of the FNN model. Dotted line shows a threshold set at 0.4. AUC: Area under the curve (**D**) Contingency matrix of the FNN model with threshold set at 0.4. PPV: positive predictive value

RA, and those differences were statistically significant (Table 1). RF and ACPA, whose titers are associated with higher incidence of RA development [39, 40], showed no difference between two groups (Table 1). Kaplan-Meier curves with optimal thresholds are shown in Supplementary Fig. 3.

Machine learning model to predict RA progression.

Although there were statistical differences in some variables between those who progressed to RA and those who did not, the differences were not markedly distinct between the two groups. Therefore, we employed ML, which can incorporate multiple variables and handle non-linear correlations, to obtain better discrimination model predicting RA progression from seronegative UA. Because MMP-3 is not measured in daily RA clinical practice, we built models from variables with or without MMP-3.

We first evaluated the performance of non-deep learning (DL) models using variables without MMP-3. ML models were built and hyperparameters were tuned using PyCaret, which automates comprehensive screening and performance comparison of ML models [41]. By comparing averaged performance measures through 5-fold cross-validation (Fig. 1A), Gradient Boosting Classifier, Random Forest Classifier, Ada Boost Classifier, Extra Trees Classifier, and Light Gradient Boosting Machine showed AUC of >0.75 (Supplementary Table 3). As the model is intended for use as a screening test in a daily clinical practice, it is important not to miss true-positive patients, and therefore sensitivity is considered the most important indicator [42, 43]. In terms of sensitivity, Random Forest Classifier and Ada Boost Classifier exceeded sensitivity of 50% (Supplementary Table 3). Adding MMP-3 improved overall performances of non-DL models (Supplementary Table 4). Random Forest Classifier, Gradient Boosting Classifier, Ada Boost Classifier, Light Gradient Boosting Machine, and Extra Trees Classifier showed AUC > 0.80. Additionally, four of them exceeded sensitivity of 60% (Supplementary Table 4), suggesting that MMP-3 is an influential variable in ML models.

Table 1 Baseline patient characteristics comparing patients who progressed to RA with those who did not in KURAMA training cohort

Baseline characteristics	Overall (*n* = 210)	RA (*n* = 57)	Non-RA (*n* = 153)	*p*-value
age (median, [Q1-Q3])	54 [44–66]	60 [49–70]	52 [42-63.5]	0.0094
sex (female%)	72.9%	63.2%	76.5%	0.079
BMI (median, [Q1-Q3])	21.6 [19.6–24]	22.8 [20.4–25]	21.2 [19.3–23.6]	0.010
Family history of RA (positive %)	28.1%	29.8%	27.5%	0.87
Smoking (current or previous %)	31.9%	42.1%	28.1%	0.077
CRP, mg/L (median, [Q1-Q3])	1 [1–6]	7 [1-19.5]	1 [0–1]	< 0.001
ESR_1h (median, [Q1-Q3])	14 [6–26]	26 [8–48]	11 [6–19]	< 0.001
RF (median, [Q1-Q3])	8 [8–8]	8 [8–8]	8 [8–8]	
ACPA (median, [Q1-Q3])	0.6 [0.6–0.6]	0.6 [0.6–0.6]	0.6 [0.6–0.6]	
MMP-3 (median, [Q1-Q3])	49.8 [32.5–96.8]	95.9 [50.8-175.25]	43.7 [30.4-66.225]	< 0.001
PhGA (median, [Q1-Q3])	13 [3–26]	25 [15.5–40.5]	8 [2–20]	< 0.001
PtGA (median, [Q1-Q3])	48.5 [21–60]	56 [40.75-75]	45.5 [18–54]	< 0.001
TJC28 (median, [Q1-Q3])	2 [0–4]	3 [1–6]	1 [0–3]	< 0.001
SJC28 (median, [Q1-Q3])	1 [0–2]	2 [1–4]	0 [0–1]	< 0.001
CDAI (median, [Q1-Q3])	9.05 [5.375–14.025]	13.75 [10.1-21.125]	7.55 [4.375–11.6]	< 0.001
HAQ-DI (median, [Q1-Q3])	0.38 [0-0.75]	0.63 [0.38-1]	0.25 [0-0.5]	< 0.001
Days to diagnosis (median, [Q1-Q3])		37 [28–76]	105 [35–427]	

Data showing descriptive statistics before imputation

Abbreviations

BMI: body mass index, CRP: C-reactive protein, ESR: erythrocyte sedimentation rate, RF: rheumatoid factor, ACPA: anti-citrullinated protein antibody, MMP-3: matrix metalloproteinase 3, PhGA: physician global assessment, PtGA: patient global assessment, TJC: 28 tender joint counts, SJC: 28 swollen joint counts, CDAI: clinical disease activity index, HAQ-DI: Health Assessment Questionnaire Disability Index

We next built a DL-based model that may outperform many classical ML approaches [25]. In this study, we modeled Feedforward Neural Network (FNN) using variables including MMP-3 (Fig. 1B). FNN's discriminatory performance on training data achieved an accuracy of 87.8%, AUC of 0.924, sensitivity of 70.6%, PPV of 75.1%, and F1 score of 0.720 after 5-fold cross-validation. The FNN exceeded all non-DL models in all performance measures.

To obtain an optimal threshold, the threshold was set at 0.4 using AUC and contingency matrix to increase sensitivity while avoiding false positives as much as possible (Fig. 1B, C, and D). Applying this threshold, the FNN model achieved sensitivity of 80.7%, PPV of 73.0%, and specificity of 88.9% in the KURAMA training cohort.

Impact of variables to the model

Despite the superior performance of the FNN model, limited explainability of the model (e.g. which variable is most important, how variables affect each other) is one of the disadvantages of DL-based models [44]. To clarify the importance of each feature in the FNN model, we computed SHAP values [44], revealing that MMP-3 had the highest impact on the model, followed by PhGA, PtGA, BMI, and age (Fig. 2A). Notably, higher values for MMP-3, PhGA, PtGA, and BMI positively contributed to the higher SHAP values (Fig. 2B and C). This means that as the values of these features increase, the model's prediction towards RA progression becomes stronger.

Model validation by external data

To ensure the robustness and generalizability of our models and to address the potential issue of overfitting, where a model performs well on the training data but fails to generalize to new, unseen data, we validated the models using an independent external cohort, the ANSWER cohort. This cohort included 140 patients with seronegative UA, of whom 45 (32.1%) progressed to RA (Supplementary Fig. 4). Median time to RA progression was 92 days (Table 2). The reason for diagnosis in RA progressors and the outcome and observation period in RA non-progressors were not available due to limited access to medical records. Significant differences were observed between RA progressors and non-progressors in CRP, ESR, MMP-3, PhGA, and HAQ-DI (Table 2).

The FNN model, which was trained and tested using the KURAMA cohort, was then applied to the ANSWER cohort, where it achieved an AUC of 0.777 (Fig. 3A). By setting the threshold at 0.4, the FNN model showed an averaged accuracy of 67.9%, sensitivity of 76.2%, specificity of 64.0%, and PPV of 50.0% in MICE-imputed datasets. A representative confusion matrix is shown in Fig. 3B. Although accuracy, specificity, and PPV decreased in the validation cohort, the model retained a reasonable level of sensitivity and AUC, demonstrating its generalizability and potential utility in clinical practice.

In summary, the FNN model demonstrated the highest predictive performance among the ML models tested in the training cohort and maintained substantial sensitivity

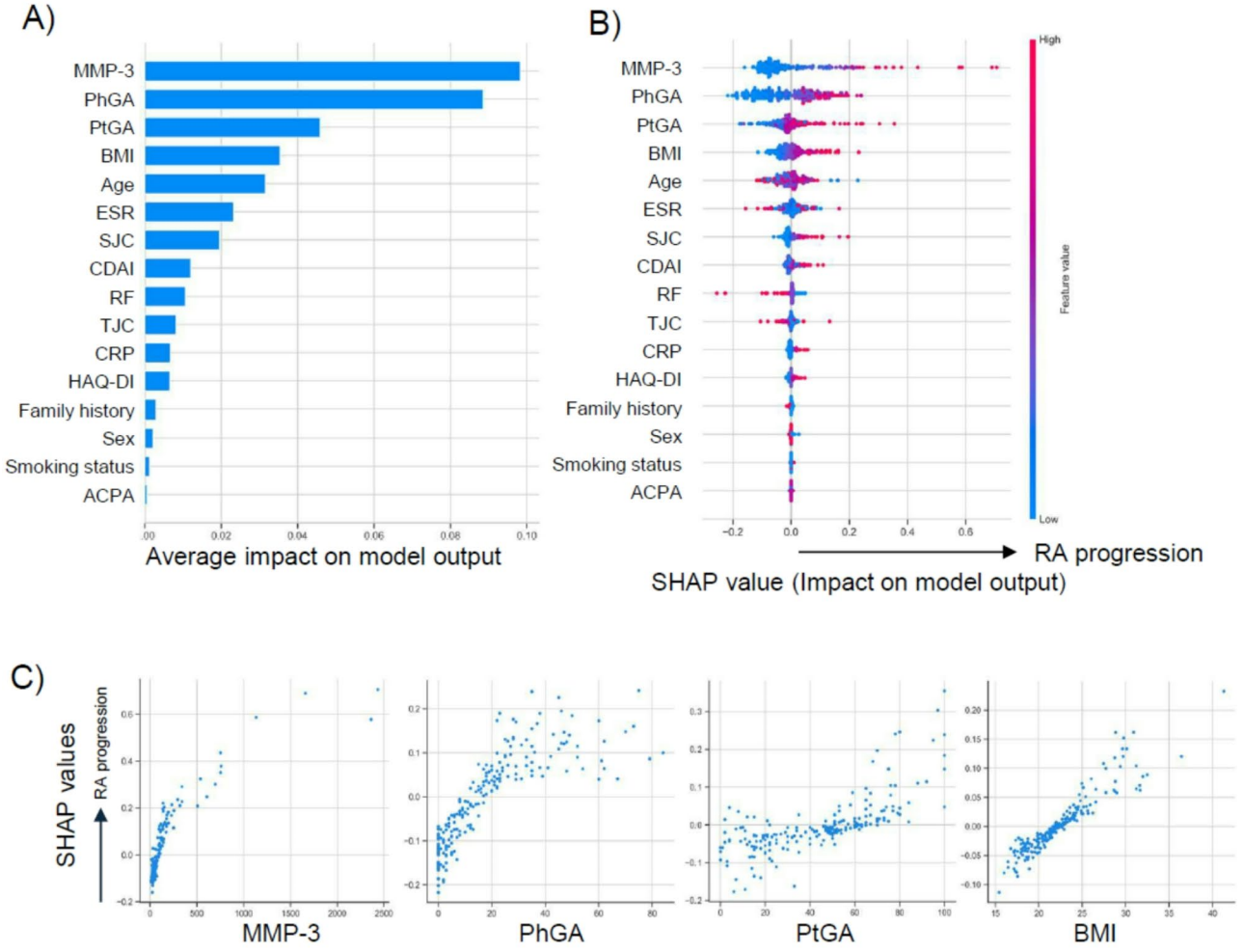

Fig. 2 Feature importance in the FNN model estimated using SHAP. (**A**) Averaged SHAP values that reflect the impact on the model. (**B**) Beeswarm plot showing the correlation between feature values and SHAP values. When the SHAP value is positive, the feature contributes positively to the prediction of RA progression. (**C**) Dependance plots showing the correlation between SHAP values and the values of the variables

in an independent validation cohort. These findings suggest that FNNs, with their ability to capture complex patterns within medical data, hold promise for predicting RA progression in patients with seronegative UA. Further research with larger and more diverse patient populations is warranted to confirm these findings and to optimize the model for practical clinical application.

Discussion

The findings of this study demonstrate the potential of ML models, particularly FNN, in predicting the progression of seronegative UA to RA. ML models, especially FNN, showed good predictive performance in two cohorts, suggesting generalizability of this FNN model.

UA's natural history varies: around 50% of cases resolve spontaneously, while about 30% progress to RA [20]. In the KURAMA and ANSWER cohorts, RA progression rates from seronegative UA were 27.1% and 32.3%, respectively. Brinkman et al. reported that patients who developed RA from UA were generally older, more often female, and exhibited higher levels of TJC, SJC, ESR, and visual analogue scale scores [22]. Another study has similarly shown that SJC and ESR are higher in patients who progress to RA [21]. In our study, older age, TJC, SJC, acute-phase reactants (CRP and ESR), and the PhGA were associated with RA progression in the KURAMA cohort, while younger age and higher CRP were significant in the ANSWER cohort. These differences illustrate the complexity of predicting RA progression using clinical parameters alone.

This study focused on seronegative UA because the sensitivity of the ACR/EULAR 2010 RA classification criteria is below 20% for seronegative RA [15, 16]. We observed in the KURAMA cohort that 27.1% of seronegative UA patients developed RA. Seropositivity in general population is associated with RA development [17, 18]. Recent studies suggest that the prevalence of seropositive and seronegative RA is becoming similar [45, 46]. The

Table 2 Baseline patient characteristics comparing patients who progressed to RA with those who did not in ANSWER validation cohort

Baseline characteristics	Overall (*n* = 140)	RA (*n* = 45)	Non-RA (*n* = 95)	*p*-value
age (median, [Q1-Q3])	62.5 [48.0–72.0]	54.0 [46.0–70.0]	64.0 [50.0–73.0]	0.112
sex (female%)	67.9%	75.6%	62.4%	0.251
BMI (median, [Q1-Q3])	21.7 [19.8–24.7]	21.6 [19.8–24.8]	21.9 [19.8–24.2]	0.617
Family history of RA (positive %)	14.3%	17.8%	12.6%	0.580
Smoking (current or previous %)	67.2%	65.0%	68.40%	0.430
CRP (median, [Q1-Q3])	2.7 [0.050–14.6]	9.4 [1.0-31.4]	2.0 [0.4–0.96]	0.011
ESR_1h (median, [Q1-Q3])	18.5 [8.0–37.0]	26.0 [12.0-44.3]	15.0 [5.0–34.0]	0.024
RF (median, [Q1-Q3])	5 [5–8]	5 [4–7]	5 [5–8]	
ACPA (median, [Q1-Q3])	0.6 [0.6–0.6]	0.6 [0.6–0.6]	0.6 [0.6–0.6]	
MMP-3 (median, [Q1-Q3])	75.7 [40.3-134.7]	95.1 [44.5-189.9]	56.9 [39.7–118.0]	0.030
PhGA (median, [Q1-Q3])	20.0 [10.5–36.5]	30.0 [12.0–50.0]	18.0 [10.0–30.0]	0.040
PtGA (median, [Q1-Q3])	50.0 [29.0-67.5]	50.0 [32.0–68.0]	50.0 [27.0–66.0]	0.371
TJC28 (median, [Q1-Q3])	2 [1–5]	3 [1–5]	2 [1–5]	0.837
SJC28 (median, [Q1-Q3])	1 [0–3]	1.0 [0–4.0]	1 [0–3]	0.943
CDAI (median, [Q1-Q3])	12.0 [8.35–15.85]	11.8 [8.68–17.6]	12.0 [8.35–15.35]	0.578
HAQ-DI (median, [Q1-Q3])	0.5 [0.13–0.88]	0.63 [0.25–1.44]	0.38 [0.13–0.75]	0.027
Days to diagnosis (median, [Q1-Q3])		92 [36–273]		

Data showing descriptive statistics before imputation

Abbreviations

BMI: body mass index, CRP: C-reactive protein, ESR: erythrocyte sedimentation rate, RF: rheumatoid factor, ACPA: anti-citrullinated protein antibody, MMP-3: matrix metalloproteinase 3, PhGA: physician global assessment, PtGA: patient global assessment, TJC: 28 tender joint counts, SJC: 28 swollen joint counts, CDAI: clinical disease activity index, HAQ-DI: Health Assessment Questionnaire Disability Index

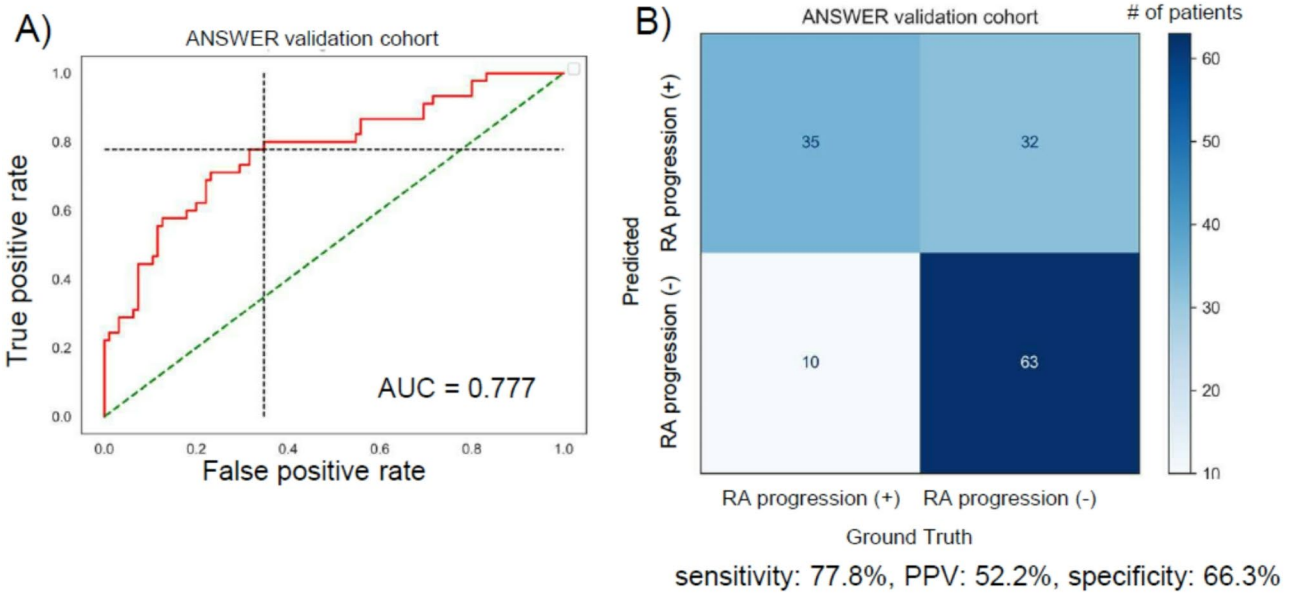

Fig. 3 FNN's Discriminatory performance in the external validation cohort. (**A**) Receiver operating characteristic curve of the FNN model applied to the ANSWER validation cohort. Dotted line shows a threshold set at 0.4. (**B**) Contingency matrix of the FNN model with threshold set at 0.4

progression rate observed in KURAMA (27.1%) indicates that seropositivity may not be a definitive predictor of RA progression.

Our ML models varied in performance, but the FNN model outperformed others in accuracy, sensitivity, PPV, and AUC. The utilization of DL or other ML models is wide spreading in medical field [25], and the previous study using a support vector machine model that incorporated clinical parameters and DNA methylation profiles in the peripheral blood mononuclear cells demonstrated more than 80% accuracy in the training cohort (*n* = 72) and 75% accuracy in the validation cohort (*n* = 8) of UA patients, including seropositive patients [21]. Our study focused solely on seronegative UA and achieved

approximately 80% sensitivity in both training and validation cohorts using clinical parameters easily obtained in practice.

One challenge with DL models is interpretability, crucial in clinical settings [47]. We addressed this by using SHAP values [44] to explain the contributions of key variables, such as MMP-3, PhGA, PtGA, BMI, and SJC. MMP-3, which had not been identified as a risk factor, had the highest impact on the model based on SHAP values. Another limitation of DL models is their computational cost. In this study, non-DL models also demonstrated fair discriminative performance in the training dataset, suggesting that non-DL models may, in some cases, be suitable for practical use.

A critical aspect of this study was the validation of our models using a multicenter cohort (ANSWER cohort). The risk of overfitting underlies ML models developed using a single cohort and may lead to a loss of generalisability [41, 47]. The FNN model trained on the KURAMA cohort maintained reasonable performance in sensitivity and AUC when externally validated on the ANSWER cohort, indicating that the model's predictive power is not confined to the training dataset. However, the FNN model demonstrated a loss of specificity in the validation dataset, reflecting the challenges and lack of consensus in diagnosing seronegative RA at an early stage. Indeed, the difference in median time to RA progression between the KURAMA training cohort and the ANSWER validation cohort differed, which could be due to multiple factors, including differences in physicians' decision thresholds and access to ultrasound and MRI. Clinical trials involving seronegative RA often employ more conservative inclusion criteria for seronegative patients, such as requiring the pre-existence of structural damage in more than three joints [48, 49]. To facilitate early diagnosis and treatment of seronegative patients before joint destruction begins, our FNN model may be of value in daily clinical practice.

This study has several limitations. First, the potential for overfitting cannot be entirely dismissed despite our validation efforts, as participants in the KURAMA and ANSWER cohorts are predominantly Asian and the sample sizes were relatively small, necessitating further validation in larger and more diverse populations. Second, the follow-up duration may not have fully captured the long-term progression from UA to RA. In terms of diagnostic challenges of seronegative RA, some patients initially diagnosed as seronegative RA may later be reclassified as other entities such as spondylarthritis or psoriatic arthritis [50]. In the ANSWER cohort, this possibility remains because of the limitation of the access to medical records, and may have contributed to the observed loss of specificity. Importantly, the proposed model does not provide diagnostic predictions in patients who were not predicted to have RA, which does not mean that follow-up of these patients is unnecessary.

In conclusion, our study highlights the potential of FNNs in predicting the progression of UA to RA, offering a noninvasive tool for predicting RA progression. Testing blood MMP-3 levels and integrating this predictive model into the clinical workflow of seronegative UA patients may improve patient outcomes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13075-025-03541-8.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

We would like to extend our heartfelt gratitude to Michie Asada, Haruo Horii, Naohiro Ito, and Masatoshi Fujii for their generous support for the KURAMA cohort.

Author contributions

T.F. conceptualized this study, analyzed data, built machine learning models, and wrote the initial draft of the manuscript. H.K. gave the expert opinion on modeling machine learning, model interpretation, and data visualization. T.F, K.Murata., A.O., K.Murakami., M.T., W.Y., K.N., A.Y., Y.E., Y.O., N.Y., H.A., T.O., Y.U., R.H., M.H., T.O., and A.M. collected and organized data. S.M. supervised the study. All authors contributed to discussion and interpretation of the results, critically reviewed the manuscript and approved the final version for submission. Manuscript guarantor: T.F.

Funding

This work was supported by Grants-in-Aid for Scientific Research (KAKENHI) (22K16359 and 24K11614). The Department of Advanced Medicine for Rheumatic Diseases, Kyoto University Graduate School of Medicine, is supported by Nagahama City, Shiga, Japan; Toyooka City, Hyogo, Japan; Asahi Kasei Pharma Corp.; and AYUMI Pharmaceutical Co. The KURAMA cohort study is supported by a grant from Daiichi Sankyo Co. Ltd. The study reported in this publication uses ANSWER Cohort supported by grants from 12 pharmaceutical companies (AbbVie GK, Asahi Kasei, Ayumi, Chugai, Eisai, Eli Lilly Japan K.K, Janssen K.K, Ono, Sanofi K.K, Taisho, Teijin Healthcare, and UCB Japan).

Data availability

Data are available on reasonable request. Data are not available without approval from the Institutional Review Board of all hospitals. Codes are available on reasonable request. Requests should be sent via email to the corresponding author

Declarations

Competing interests

T.F. has received speaker fees from AbbVie GK, Astellas Pharma Inc, Asahi Kasei Pharma Corp., Chugai Pharmaceutical Co., Ltd., Eisai Co., Ltd., Daiichi Sankyo Co. Ltd., Mitsubishi Tanabe Pharma Co., Taisho Pharmaceutical Co., Ltd., and Janssen Pharmaceutical K.K. K.Murata received research grants and/or speaker fees from Eisai Co., Ltd., Chugai Pharmaceutical Co., Ltd., Pfizer Inc., Bristol-Myers Squibb, Mitsubishi Tanabe Pharma Corporation, UCB Japan Co., Ltd., Daiichi Sankyo Co., Ltd., and Astellas Pharma Inc. A.O. received research grants and speaker fees from Pfizer Inc., Bristol Myers Squibb, Advantest, Asahi Kasei Pharma Corp., Chugai Pharmaceutical Co. Ltd., Eli Lilly Japan K. K., Ono Pharmaceutical Co., UCB Japan Co., Mitsubishi Tanabe Pharma Co., Eisai Co. Ltd., AbbVie Inc., Takeda Pharmaceutical Co. Ltd., and Daiichi Sankyo Co. Ltd. K.Murakami received speaking and/or consulting fees from AbbVie GK, Eisai Co., Ltd., Pfizer Inc., Chugai Pharmaceutical Co., Ltd., Mitsubishi Tanabe Pharma Corp., Bristol-Myers Squibb, Daiichi Sankyo Co., Ltd., Janssen Pharmaceutical

K.K., and Asahi Kasei Pharma Corp. M.T. has received research grants and/or speaker fees from AbbVie GK, Asahi Kasei Pharma Corp., Astellas Pharma Inc., Chugai Pharmaceutical Co., Ltd., Daiichi Sankyo Co., Ltd., Eisai Co., Ltd., Eli Lilly Japan K.K., Janssen Pharmaceutical K.K., Kyowa Kirin Co., Ltd., Pfizer Inc., Taisho Pharmaceutical Co., Ltd., Tanabe Mitsubishi Pharma Corp., Teijin Pharma, Ltd., UCB Japan Co., Ltd. K.N. has received speaker fees from Pfizer Inc., Asahi Kasei Pharma Corp., Chugai Pharmaceutical Co. Ltd., Eli Lilly Japan K. K., Ono Pharmaceutical Co., UCB Japan Co., Mitsubishi Tanabe Pharma Co., Eisai Co. Ltd., AbbVie Inc., Janssen Pharmaceutical K.K., Taisho Pharmaceutical Co., Ltd, and Viatris Inc. Y.E. received research grants and/or speaker fees from Asahi Kasei, Eisai, Eli Lilly, Mitsubishi Tanabe, Ono Pharmaceuticals, and Taisho. Y.E. is affiliated with the Department of Sports Medical Biomechanics, Osaka University Graduate School of Medicine, supported by Asahi Kasei. Y.O. has received lecture fees from Chugai, Pfizer and Ono. H. A. has received speaker fees from Asahi Kasei Pharma, Chugai Pharmaceutical Co., Ltd, Eisai Co., Ltd, Taisho Pharmaceutical Co., and Mitsubishi Tanabe Pharma. T.O. has received research grants from Abbvie, Asahi Kasei, Chugai, Eisai, and Tanabe Mitsubishi, and speaker fees from Abbvie, Asahi Kasei, Chugai, Eisai, Eli Lilly, Janssen, Novartis Pharma and Tanabe Mitsubishi. K. Y.U. has received speaker fees from AbbVie, Asahi Kasei Pharma, Astellas, Bristol Meyers Squibb, Chugai, Eisai, Eli Lilly, Gilead Sciences, and Taiho Pharmaceutical. R.H. received speaker fees from AbbVie, Eli Lilly, Eisai, and Asahi Kasei. M.H. received research grants and/or speaker fees from AbbVie, Asahi Kasei, Astellas, Brystol Meyers, Chugai, EA Pharma, Eisai, Daiichi Sankyo, Eli Lilly, Novartis Pharma, Taisho Toyama, and Tanabe Mitsubishi. A.M. received honorarium and research grants from AbbVie G.K., Chugai Pharmaceutical Co. Ltd., Eli Lilly Japan K.K., Eisai Co. Ltd., Pfizer Inc., Bristol-Myers Squibb., Mitsubishi Tanabe Pharma Co., Astellas Pharma Inc., Asahi Kasei Pharma Corp., and Gilead Sciences Japan. S.M. received research grants and/or speaker fees from Takeda, Eisai, Asahi-Kasei, Astellas, Pfizer, Taisho, Mitsubishi-Tanabe, and Chugai. Other authors have no conflicts of interest to disclose concerning this manuscript.

Ethics approval and consent to participate

The institutional review board at all hospitals participating in the KURAMA and the ANSWER cohorts approved this study. Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Author details

[1]Department of Advanced Medicine for Rheumatic Diseases, Graduate School of Medicine, Kyoto University, 54 Shogoin Kawaharacho, Sakyo, Kyoto, Japan
[2]Department of Orthopaedic Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[3]Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[4]Center for Cancer Immunotherapy and Immunobiology, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[5]Department of Health Information Management, Kurashiki Sweet Hospital, Kurashiki, Japan
[6]Department of Internal Medicine (IV), Osaka Medical and Pharmaceutical University, Takatsuki, Japan
[7]Department of Sports Medical Biomechanics, Osaka University Graduate School of Medicine, Suita, Japan
[8]Department of Respiratory Medicine and Clinical Immunology, Osaka University, Suita, Japan
[9]First Department of Internal Medicine, Kansai Medical University, Hirakata, Japan
[10]Center for Senile Degenerative Disorders (CSDD), Osaka Metropolitan University Graduate School of Medicine, Osaka, Japan
[11]Department of Rheumatology and Clinical Immunology, Kobe University Graduate School of Medicine, Kobe, Japan
[12]Department of Clinical Laboratory, Kobe University Hospital, Kobe, Japan
[13]Department of Orthopaedic Surgery, Nara Medical University, Kashihara, Japan
[14]Department of Clinical Immunology, Osaka Metropolitan University, Osaka, Japan
[15]Department of Rheumatology and Clinical Immunology, Kyoto University, Kyoto, Japan

Received: 3 February 2025 / Accepted: 19 March 2025
Published online: 26 March 2025

References

1. Smolen JS, Aletaha D, Barton A, et al. Rheumatoid arthritis. Nat Rev Dis Primers. 2018;4:18001.
2. Myasoedova E, Matteson EL. Updates on interstitial lung disease and other selected extra-articular manifestations of rheumatoid arthritis. Curr Opin Rheumatol. 2024;36(3):203–8.
3. Sorensen J, Hetland ML. All departments of rheumatology in D. Diagnostic delay in patients with rheumatoid arthritis, psoriatic arthritis and ankylosing spondylitis: results from the Danish nationwide DANBIO registry. Ann Rheum Dis. 2015;74(3):e12.
4. Monti S, Montecucco C, Bugatti S, Caporali R. Rheumatoid arthritis treatment: the earlier the better to prevent joint damage. RMD Open. 2015;1(Suppl 1):e000057.
5. 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(9):2569–81.
6. Norli ES, Brinkmann GH, Kvien TK, et al. Diagnostic spectrum and 2-year outcome in a cohort of patients with very early arthritis. RMD Open. 2017;3(2):e000573.
7. Hazes JM, Luime JJ. The epidemiology of early inflammatory arthritis. Nat Rev Rheumatol. 2011;7(7):381–90.
8. Novella-Navarro M, Plasencia-Rodriguez C, Nuno L, Balsa A. Risk factors for developing rheumatoid arthritis in patients with undifferentiated arthritis and inflammatory arthralgia. Front Med (Lausanne). 2021;8:668898.
9. van Steenbergen HW, da Silva JAP, Huizinga TWJ, van der Helm-van Mil AHM. Preventing progression from arthralgia to arthritis: targeting the right patients. Nat Rev Rheumatol. 2018;14(1):32–41.
10. Wevers-de Boer KV, Heimans L, Huizinga TW, Allaart CF. Drug therapy in undifferentiated arthritis: a systematic literature review. Ann Rheum Dis. 2013;72(9):1436–44.
11. Verstappen M, Matthijssen XME, van der Helm-van Mil AHM. Undifferentiated arthritis: a changing population who did not benefit from enhanced disease-modifying anti-rheumatic drug strategies-results from a 25 year longitudinal inception cohort. Rheumatology (Oxford). 2022;61(8):3212–22.
12. van der Helm AH, le Cessie S, van Dongen H, Breedveld FC, Toes RE, Huizinga TW. A prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: how to guide individual treatment decisions. Arthritis Rheum. 2007;56(2):433–40.
13. Ghosh K, Chatterjee A, Ghosh S, et al. Validation of Leiden score in predicting progression of rheumatoid arthritis in undifferentiated arthritis in Indian population. Ann Med Health Sci Res. 2016;6(4):205–10.
14. Aletaha D, Smolen JS. Diagnosis and management of rheumatoid arthritis: A review. JAMA. 2018;320(13):1360–72.
15. Kaneko Y, Kuwana M, Kameda H, Takeuchi T. Sensitivity and specificity of 2010 rheumatoid arthritis classification criteria. Rheumatology (Oxford). 2011;50(7):1268–74.
16. Biliavska I, Stamm TA, Martinez-Avila J, et al. Application of the 2010 ACR/EULAR classification criteria in patients with very early inflammatory arthritis: analysis of sensitivity, specificity and predictive values in the SAVE study cohort. Ann Rheum Dis. 2013;72(8):1335–41.
17. Bos WH, Wolbink GJ, Boers M, et al. Arthritis development in patients with arthralgia is strongly associated with anti-citrullinated protein antibody status: a prospective cohort study. Ann Rheum Dis. 2010;69(3):490–4.
18. Nielsen SF, Bojesen SE, Schnohr P, Nordestgaard BG. Elevated rheumatoid factor and long term risk of rheumatoid arthritis: a prospective cohort study. BMJ. 2012;345:e5244.
19. Rech J, Tascilar K, Hagen M, et al. Abatacept inhibits inflammation and onset of rheumatoid arthritis in individuals at high risk (ARIAA): a randomised, international, multicentre, double-blind, placebo-controlled trial. Lancet. 2024;403(10429):850–9.
20. van Aken J, van Dongen H, le Cessie S, Allaart CF, Breedveld FC, Huizinga TW. Comparison of long term outcome of patients with rheumatoid arthritis presenting with undifferentiated arthritis or with rheumatoid arthritis: an observational cohort study. Ann Rheum Dis. 2006;65(1):20–5.
21. de la Calle-Fabregat C, Niemantsverdriet E, Canete JD, et al. Prediction of the progression of undifferentiated arthritis to rheumatoid arthritis using DNA methylation profiling. Arthritis Rheumatol. 2021;73(12):2229–39.

22. Brinkmann GH, Norli ES, Kvien TK, et al. Disease characteristics and rheumatoid arthritis development in patients with early undifferentiated arthritis: A 2-year followup study. J Rheumatol. 2017;44(2):154–61.
23. Habehh H, Gohel S. Machine learning in healthcare. Curr Genomics. 2021;22(4):291–300.
24. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun. 2020;11(1):3923.
25. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.
26. Baker JF. Diagnosis and differential diagnosis of rheumatoid arthritis. In: UpToDate, ed. UpToDate: Wolters Kluwer. (Accessed on July 7, 2024.).
27. van der Helm AH, Huizinga TW. The 2010 ACR/EULAR criteria for rheumatoid arthritis: do they affect the classification or diagnosis of rheumatoid arthritis? Ann Rheum Dis. 2012;71(10):1596–8.
28. Fujii T, Murata K, Onizawa H, et al. Management and treatment outcomes of rheumatoid arthritis in the era of biologic and targeted synthetic therapies: evaluation of 10-year data from the KURAMA cohort. Arthritis Res Ther. 2024;26(1):16.
29. Fujii T, Murata K, Onizawa H, et al. The influence of medial cuneiform inclination on postoperative hallux valgus recurrence in rheumatoid arthritis patients: insights from the KURAMA cohort study. Int J Rheum Dis. 2024;27(5):e15168.
30. Iwasaki T, Watanabe R, Ito H, et al. Dynamics of type I and type II interferon signature determines responsiveness to Anti-TNF therapy in rheumatoid arthritis. Front Immunol. 2022;13:901437.
31. Murata K, Uozumi R, Hashimoto M, et al. The real-world effectiveness of anti-RANKL antibody denosumab on the clinical fracture prevention in patients with rheumatoid arthritis: the ANSWER cohort study. Mod Rheumatol. 2022;32(4):834–8.
32. Hayashi S, Tachibana S, Maeda T et al. Real-world comparative study of the efficacy of Janus kinase inhibitors in patients with rheumatoid arthritis: the ANSWER cohort study. Rheumatology (Oxford) 2023.
33. Ebina K, Etani Y, Maeda Y et al. Drug retention of biologics and Janus kinase inhibitors in patients with rheumatoid arthritis: the ANSWER cohort study. RMD Open 2023; 9(3).
34. Nakayama Y, Onishi A, Yamamoto W, et al. Safety of Janus kinase inhibitors compared to biological DMARDs in patients with rheumatoid arthritis and renal impairment: the ANSWER cohort study. Clin Exp Med. 2024;24(1):97.
35. Aletaha D, Nell VP, Stamm T, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther. 2005;7(4):R796–806.
36. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23(2):137–45.
37. Rajput D, Wang WJ, Chen CC. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics. 2023;24(1):48.
38. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55–63.
39. del Puente A, Knowler WC, Pettitt DJ, Bennett PH. The incidence of rheumatoid arthritis is predicted by rheumatoid factor titer in a longitudinal population study. Arthritis Rheum. 1988;31(10):1239–44.
40. Bizzaro N, Bartoloni E, Morozzi G, et al. Anti-cyclic citrullinated peptide antibody titer predicts time to rheumatoid arthritis onset in patients with undifferentiated arthritis: results from a 2-year prospective study. Arthritis Res Ther. 2013;15(1):R16.
41. Saito R, Fujii T, Murata K, et al. Prediction models incorporating second metacarpal cortical index for osteoporosis in rheumatoid arthritis: externally validated machine learning models developed using data from the KURAMA cohort. Int J Rheum Dis. 2024;27(10):e15358.
42. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Educ Anaesth Crit Care Pain. 2008;8(6):221–3.
43. Vetter TR, Schober P, Mascha EJ. Diagnostic testing and Decision-Making: beauty is not just in the eye of the beholder. Anesth Analg. 2018;127(4):1085–91.
44. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749–60.
45. Myasoedova E, Davis J, Matteson EL, Crowson CS. Is the epidemiology of rheumatoid arthritis changing? Results from a population-based incidence study, 1985–2014. Ann Rheum Dis. 2020;79(4):440–4.
46. Krajewska-Wlodarczyk M, Szelag M, Batko B, et al. Rheumatoid arthritis epidemiology: a nationwide study in Poland. Rheumatol Int. 2024;44(6):1155–63.
47. Huang D, Cogill S, Hsia RY, Yang S, Kim D. Development and external validation of a pretrained deep learning model for the prediction of non-accidental trauma. NPJ Digit Med. 2023;6(1):131.
48. Combe B, Kivitz A, Tanaka Y, et al. Filgotinib versus placebo or adalimumab in patients with rheumatoid arthritis and inadequate response to methotrexate: a phase III randomised clinical trial. Ann Rheum Dis. 2021;80(7):848–58.
49. Taylor PC, Keystone EC, van der Heijde D, et al. Baricitinib versus placebo or adalimumab in rheumatoid arthritis. N Engl J Med. 2017;376(7):652–62.
50. Perniola S, Chimenti MS, Spinelli FR et al. Rheumatoid Arthritis from Easy to Complex Disease: From the 2022 GISEA International Symposium. *J Clin Med* 2023; 12(8).

Publisher's note