Journal of
Translational Medicine

RESEARCH Open Access

Predicting higher risk factors for COVID-19 short-term reinfection in patients with rheumatic diseases: a modeling study based on XGBoost algorithm

Yao Liang[1†], Siwei Xie[2†], Xuqi Zheng[1], Xinyu Wu[1], Sijin Du[3] and Yutong Jiang[1*]

Abstract

Background Corona virus disease 2019 (COVID-19) reinfection, particularly short-term reinfection, poses challenges to the management of rheumatic diseases and may increase adverse clinical outcomes. This study aims to develop machine learning models to predict and identify the risk of short-term COVID-19 reinfection in patients with rheumatic diseases.

Methods We developed four prediction models using explainable machine learning to assess the risk of short-term COVID-19 reinfection in 543 patients with rheumatic diseases. Psychological health was evaluated using the Functional Assessment of Chronic Illness Therapy Fatigue (FACIT-F) scale, the Patient Health Questionnaire-9 (PHQ-9), the Generalized Anxiety Disorder 7-item (GAD-7) questionnaire, and the Pittsburgh Sleep Quality Index (PSQI) scale. Health status and disease activity were assessed using the EuroQol-5 Dimension-3 Level (EQ-5D-3L) descriptive system and the Visual Analogue Score (VAS) scale. The model performance was assessed by Area Under the Receiver Operating Characteristic Curve (AUC), Area Under the Precision-Recall Curve (AUPRC), and the geometric mean of sensitivity and specificity (G-mean). SHapley Additive exPlanations (SHAP) analysis was used to interpret the contribution of each predictor to the model outcomes.

Results The eXtreme Gradient Boosting (XGBoost) model demonstrated superior performance with an AUC of 0.91 (95% CI 0.87–0.95). Significant factors of short-term reinfection included glucocorticoid taper (OR = 2.61, 95% CI 1.38–4.92), conventional synthetic disease-modifying antirheumatic drugs (csDMARDs) taper (OR = 2.97, 95% CI 1.90–4.64), the number of symptoms (OR = 1.24, 95% CI 1.08–1.42), and GAD-7 scores (OR = 1.07, 95% CI 1.02–1.13). FACIT-F scores were associated with a lower likelihood of short-term reinfection (OR = 0.95, 95% CI 0.93–0.96). Besides, we found that the GAD-7 score was one of the most important predictors.

Conclusion We developed explainable machine learning models to predict the risk of short-term COVID-19 reinfection in patients with rheumatic diseases. SHAP analysis highlighted the importance of clinical and psychological factors. Factors included anxiety, fatigue, depression, poor sleep quality, high disease activity during initial infection, and the use of glucocorticoid taper were significant predictors. These findings underscore the need for targeted preventive measures in this patient population.

[†] Yao Liang and Siwei Xie have contributed equally to this work.

*Correspondence:
Yutong Jiang
Jiangyt7@mail.sysu.edu.cn
Full list of author information is available at the end of the article

Keywords COVID-19, Short-term reinfection, Rheumatism, Machine learning, XGBoost, SHAP analysis, Psychological factors

Introduction

It has been proven that even with protective antibodies after COVID-19 infection, reinfection can occur [1, 2]. In the context of emerging variants, the overall prevalence of COVID-19 reinfection increases significantly to 11.94–28.3% [3–5]. It was believed that the weakening or decline of the humoral immune response over time is the reason for reinfection [1, 2]. For patients with rheumatic diseases, due to their immune dysregulation, they were more susceptible to reinfection and had a shorter interval between two consecutive infections than the general population [6].

Given that COVID-19 reinfection, especially short-term reinfection, not only increased the risk of long COVID [7], but also imposed a greater overall burden of disease on rheumatic patients. It is of profound significance to construct a model for predicting the risk of COVID-19 short-term reinfection. In this study, COVID-19 short-term reinfection was defined as reoccurrence of positive COVID-19 test results for successive two times within 6 months after all the COVID-19 related symptoms were alleviated. Machine learning and fixed effect logistic regression were employed to build short-term reinfection prediction models for patients with rheumatic diseases and identify risk factors.

Methods

Study population

This study included patients confirmed with rheumatic disorders who visited the rheumatology department of our hospital from October 2021 to August 2023 and had confirmed COVID-19 infection with positive result of RT-PCR (reverse transcriptase-polymerase chain reaction) or AgPOCTs (antigen point-of-care tests) for COVID-19. Participants were informed of the objectives in this study and gave informed written consent. Patients who fit any of the following points were excluded: (1) Not diagnosed with rheumatism; (2) Suffer from serious mental disease; (3) Being unable to understand used questionnaires; (4) Having missing data; (5) Having noisy values. After applying the exclusion criteria, out of 769 patients visited the rheumatology department, 642 cases were included in the study. In the preprocessing steps, 99 patient record values were removed, leaving 543 cases, including 64 patients with short-term reinfection and 479 patients without short-term reinfection.

Outcomes

The primary outcome was whether patients with rheumatic diseases got short-term COVID-19 reinfection (within 6 months) as categorical variables.

Feature selection

A number of demographic and clinical variables were collected (sex, age, vaccination status, date of previous infection, date of reinfection or last follow-up, the number of rheumatic diseases and medication usage of glucocorticoid, csDMARDs, biological disease-modifying antirheumatic drugs (bDMARDs) and targeted synthetic disease-modifying antirheumatic drugs (tsDMARDs). In addition, four psychometric measures and two health measures were applied. Four psychometric measures were the FACIT-F scale [8–10], the PHQ-9 [11], the GAD-7 questionnaire [12] and the PSQI scale [13], respectively. Two health measures were EQ-5D-3L descriptive system [14, 15] and VAS scale [16], respectively. VAS scale was patients' self-report outcome which helped rapidly assess patient perspectives of rheumatic disease activity. These scales have shown good validity and reliability, and have been used widely among Chinese patients with chronic diseases. Each patient filled out these questionnaires under the guidance of a physician.

Machine learning model development

Predictor variables

The predictor variables selected for inclusion in the model were based on their relevance to the disease control of rheumatic disease and the pathophysiology, immunology, and social psychology of COVID-19. We totally selected 68 predictor variables, and detailed all these features in the Supplementary Table 1, including their meanings.

Model performance and evaluation

The dataset will firstly leave 20% hold-out data that do not include in any model training part. For the remaining 80% data, we split it into training (80%) and testing sets (20%), combing with cross-validation. A comparative analysis of four machine learning algorithms was conducted on the training subset: Light Gradient Boosting Machine (LightGBM) [17], XGBoost [18], Random Forest [19], and Lasso and Elastic-Net Regularized Generalized Linear Models (Glmnet). [20] For the hyperparameter tuning, we used five-fold cross-validations [21] combined with 15 times grid search [22]. The Models performance were assessed by three main metrics: AUC [23], AUPRC [24], baseline AUPRC, and the G-mean [25].

SHAP feature importance analysis

To better interpret and understand the influence of each feature on the prediction outcomes, we computed the SHAP values [26]. The SHAP analysis is a game theory-based approach for interpreting machine learning models, offering a unified measure of feature importance. We chose the XGBoost machine learning algorithm to conduct the SHAP analysis as it has superior performance and ability to handle high-dimensional or complex patterns data effectively. The SHAP values would show the contribution of each feature to the outcome, including: (1) Beeswarm plots with both the positive and negative directions; (2) Bar plots with the absolute SHAP value.

Statistical analysis

Descriptive statistics were conducted in the total cohort and stratified by two subgroups: (1) patients without short-term Covid-19 reinfection (within 6 months); (2) patients with short-term Covid-19 reinfection (within 6 months).

Primary analysis

The primary analysis aimed to identify patient and clinical characteristics associated with the likelihood of short-term COVID-19 reinfection (within 6 months) among patients with rheumatic diseases. Initially, we employed machine learning models to identify potential risk factors. SHAP values were then calculated to interpret the contributions of each feature to the model predictions. Based on these results, key risk factors were selected for the fixed effect logistic regression models [27], controlling the interval between initial COVID-19 infection and reinfection as a fixed effect variable, which can adjust the confounding of temporal variance.

This approach allowed for robust estimation of associations between the selected covariates and the primary outcome. We calculated the odds ratios (ORs) with 95% confidence intervals (CIs), p values, p-adjust values, and e-values to quantify and assess the significance of these associations. P-adjust values used the benjamini-Hochberg (BH) method to control the false discovery rate [28], and e-values assessed the robustness of the observed associations to potential unmeasured confounding [29]. All statistical tests were two-sided, with a nominal type I error rate of $\alpha = 0.05$ indicating statistical significance.

Secondary analysis

Secondary analyses included survival analysis and generalized additive models (GAMs) [30] to further explore the data. Kaplan–Meier survival analysis [31] was performed to assess the time to reinfection, incorporating right-censored data to provide more comprehensive insights into the duration between initial infection and reinfection. This method was chosen to account for time-to-event data and to handle the variable follow-up times among patients. Additionally, GAMs were employed to analyze continuous variables, allowing for the modeling of potential non-linear relationships between covariates and the primary outcome (likelihood of short-term COVID-19 reinfection (within 6 months)). The use of GAMs facilitated a flexible approach to capture complex interactions and patterns in the data, thereby enhancing the accuracy and robustness of the findings.

All analyses were performed by the R software, version 4.3.2 (R Project for Statistical Computing). Additional analysis details are provided in the Supplementary Methods.

Results

The demographic, clinical and medication characteristics of 543 patients were presented in Table 1. More than half of patients were females (69.43%). The mean age of patients with short-term reinfection was 37.27 (12.34) years old, while the mean age of patients without short-term reinfection was 40.16 (14.53). Among the age groups, the majority of patients were from 20 to 34 years old (40.88%), followed by 35–44 (22.10%), 45–54 (16.94%), 55–64 (10.87%), more than 65 (6.08%), and less than 19 years old (3.13%). Systemic lupus erythematosus (SLE) (23.76%), rheumatoid arthritis (RA) (22.65%) and spondyloarthritis (SpA) (17.49%) were three most common rheumatic diseases. The majority of participants have had the diagnosed with single rheumatic disease (87.11%). The csDMARDs were used in 25.50% overall patients, 25.89% patients without short-term reinfection, and in 18.75% patients with short-term reinfection. The bDMARDs and tsDMARDs were used in 20.44% overall patients, 19.62% patients without short-term reinfection, and in 26.56% patients with short-term reinfection. The mean FACIT-F was 37.69 (9.92) in overall patients and 33.36 (10.99) in patients with short-term reinfection. The mean PHQ-9 was 5.92 (6.05) in overall patients and 7.89 (6.87) in patients with short-term reinfection. Patients with short-term reinfection also had highest mean GAD-7 value and PSQI value. For the health status score, during the initial covid-19 infection, the mean patient self-report outcome was 4.07 (2.82), while the second time (currently) patient self-report outcome was 7.03 (1.74). Figure 1 showed the overall study design and modeling analysis steps.

Table 1 Demographic and clinical characteristics of patients

	Total patients (N = 543)	**Patients without short-term reinfection (N = 479)**	**Patients with short-term reinfection (N = 64)**
Demographics			
Sex, n (%)			
Male	166 (30.57)	148 (30.90)	18 (28.12)
Female	377 (69.43)	331 (69.10)	46 (71.88)
Age, years			
Mean (sd)	39.82 (14.31)	40.16 (14.53)	37.27 (12.34)
Age group, years, n (%)			
≤ 19	17 (3.13)	16 (3.34)	1 (1.56)
20–34	222 (40.88)	192 (40.08)	30 (46.88)
35–44	120 (22.10)	102 (21.29)	18 (28.12)
45–54	92 (16.94)	84 (17.54)	8 (12.50)
55–64	59 (10.87)	55 (11.48)	4 (6.25)
≥ 65	33 (6.08)	30 (6.27)	3 (4.69)
Rheumatism			
Number of rheumatism, n (%)			
1	473 (87.11)	421 (87.89)	52 (81.25)
2	55 (10.13)	46 (9.60)	9 (14.06)
3	11 (2.02)	9 (1.88)	2 (3.13)
4	4 (0.74)	3 (0.63)	1 (1.56)
RA, n (%)	123 (22.65)	110 (22.96)	13 (20.31)
SpA, n (%)	95 (17.49)	78 (16.28)	17 (26.56)
SLE, n (%)	129 (23.76)	116 (24.22)	13 (20.31)
pSS, n (%)	58 (10.68)	51 (10.65)	7 (10.94)
APS, n (%)	5 (0.92)	3 (0.63)	2 (3.13)
SSc, n (%)	18 (3.31)	16 (3.34)	2 (3.13)
PM/DM, n (%)	33 (6.08)	28 (5.85)	5 (7.81)
IgG4-RD, n (%)	3 (0.55)	2 (0.42)	1 (1.56)
Vasculitis, n (%)	15 (2.76)	13 (2.71)	2 (3.13)
AOSD, n (%)	6 (1.10)	6 (1.25)	n/a
BD, n (%)	11 (2.03)	10 (2.09)	1 (1.56)
RPC, n (%)	1 (0.18)	1 (0.21)	n/a
Gout/HUA, n (%)	41 (7.55)	36 (7.52)	5 (7.81)
OA, n (%)	18 (3.31)	16 (3.34)	2 (3.13)
FMS, n (%)	4 (0.74)	3 (0.63)	1 (1.56)
Rheumatic fever, n (%)	7 (1.29)	4 (0.84)	3 (4.69)
CTD, n (%)	52 (9.58)	48 (10.02)	4 (6.25)
Others, n (%)	13 (2.39)	11 (2.30)	2 (3.13)
csDMARDs			
Number of csDMARDs, n (%)			
1	136 (25.05)	124 (25.89)	12 (18.75)
2	118 (21.73)	103 (21.50)	15 (23.44)
3	25 (4.60)	21 (4.38)	4 (6.25)
4	9 (1.66)	8 (1.67)	1 (1.56)
5	3 (0.55)	3 (0.63)	n/a
Treated with csDMARDs, n (%)	312 (57.46)	275 (57.41)	37 (57.81)
HCQ, n (%)	161 (29.65)	142 (29.64)	19 (29.69)
MTX, n (%)	92 (16.94)	80 (16.70)	12 (18.75)
LEF, n (%)	18 (3.31)	16 (3.34)	2 (3.13)

Table 1 (continued)

	Total patients (N = 543)	Patients without short-term reinfection (N = 479)	Patients with short-term reinfection (N = 64)
MMF, n (%)	70 (12.89)	64 (13.36)	6 (9.38)
CTX, n (%)	10 (1.84)	9 (1.88)	1 (1.56)
CsA, n (%)	27 (4.97)	23 (4.80)	4 (6.25)
Tacrolimus, n (%)	11 (2.03)	11 (2.30)	n/a
Iguratimod, n (%)	34 (6.26)	28 (5.85)	6 (9.38)
SASP, n (%)	6 (1.10)	5 (1.04)	1 (1.56)
Tripterygium wilfordii Hook F, n (%)	12 (2.21)	11 (2.30)	1 (1.56)
Total glucosides of paeony, n (%)	14 (2.58)	13 (2.71)	1 (1.56)
AZA, n (%)	13 (2.39)	12 (2.51)	1 (1.56)
bDMARDs and tsDMARDs			
Number of bDMARDs and tsDMARDs, n (%)			
1	111 (20.44)	94 (19.62)	17 (26.56)
2	3 (0.55)	1 (0.21)	2 (3.13)
Treated with bDMARDs and tsDMARDs, n (%)	114 (20.99)	95 (19.83)	19 (29.69)
TNF receptor-IgG fusion protein, n (%)	31 (5.71)	27 (5.64)	4 (6.25)
Infliximab, n (%)	3 (0.55)	3 (0.63)	n/a
Adalimumab, n (%)	21 (3.87)	15 (3.13)	6 (9.38)
Golimumab, n (%)	1 (0.18)	1 (0.21)	n/a
Cetocilizumab, n (%)	5 (0.92)	1 (0.21)	4 (6.25)
Secukinumab, n (%)	24 (4.42)	19 (3.97)	5 (7.81)
Tocilizumab, n (%)	5 (0.92)	5 (1.04)	n/a
Belimumab, n (%)	7 (1.29)	7 (1.46)	n/a
Telitacicept, n (%)	7 (1.29)	6 (1.25)	1 (1.56)
Abatacept, n (%)	1 (0.18)	1 (0.21)	n/a
RTX, n (%)	1 (0.18)	1 (0.21)	n/a
Tofacitinib, n (%)	18 (3.31)	16 (3.34)	2 (3.13)
Baricitinib, n (%)	12 (2.21)	10 (2.09)	2 (3.13)
Other bDMARDs and tsDMARDs, n (%)	11 (2.03)	10 (2.09)	1 (1.56)
Drug adjustment			
Glucocorticoid taper, n (%)	96 (17.68)	82 (17.12)	14 (21.88)
csDMARDs taper, n (%)	133 (24.49)	114 (23.80)	19 (29.69)
bDMARDs and tsDMARDs taper, n (%)	87 (16.02)	73 (15.24)	14 (21.88)
Vaccinated (COVID-19), n (%)	408 (75.14)	357 (74.53)	51 (79.69)
Adverse reactions			
Adverse reactions (after vaccination), n (%)			
1	108 (19.89)	97 (20.25)	11 (17.19)
2	44 (8.10)	38 (7.93)	6 (9.38)
3	20 (3.68)	15 (3.13)	5 (7.81)
4	11 (2.03)	11 (2.30)	n/a
5	5 (0.92)	4 (0.84)	1 (1.56)
6	3 (0.55)	2 (0.42)	1 (1.56)
Number of symptoms, n (%)			
1–5	299 (55.06)	276 (57.62)	23 (35.94)
6–10	204 (37.57)	175 (36.53)	29 (45.31)
> 10	33 (6.08)	21 (4.38)	12 (18.75)
Number of residual symptoms, n (%)			
1	78 (14.36)	70 (12.89)	8 (12.50)
2	34 (6.26)	25 (4.60)	9 (14.06)

Table 1 (continued)

	Total patients (N = 543)	Patients without short-term reinfection (N = 479)	Patients with short-term reinfection (N = 64)
3	21 (3.87)	16 (2.95)	5 (7.81)
≥ 4	21 (3.87)	15 (2.76)	6 (9.38)
Influenza virus infection (after), n (%)	22 (4.05)	14 (2.92)	8 (12.50)
Psychological health surveys			
FACIT-F			
Mean (sd)	37.69 (9.92)	38.27 (9.63)	33.36 (10.99)
PHQ-9			
Mean (sd)	5.92 (6.05)	5.66 (5.89)	7.89 (6.87)
GAD-7			
Mean (sd)	4.31 (5.36)	4.02 (5.13)	6.48 (6.50)
PSQI			
Mean (sd)	6.91 (3.63)	6.78 (3.66)	7.87 (3.21)
Health status score			
Patient self-report outcome (previous infection)			
Mean (sd)	4.07 (2.82)	3.98 (2.83)	4.77 (2.69)
Patient self-report outcome (current)			
Mean (sd)	7.03 (1.74)	7.10 (1.71)	6.48 (1.86)
EQ-5D-3L index value			
Mean (sd)	0.94 (0.10)	0.94 (0.10)	0.93 (0.08)

Performance of the machine learning models

Among the four machine learning models: LightGBM, XGBoost, random forest, and glmnet, we found that the XGBoost model showed better performance in terms of distinguishing between patients who would experience short-term COVID-19 reinfection and those who would not (AUC = 0.91, 95% CI 0.87–0.95) (Table 2; Fig. 2). The random forest model (AUC = 0.89, 95% CI 0.85–0.93) and LightGBM model (AUC = 0.88, 95% CI 0.83–0.92) followed. For the AUPRC, the random forest model indicated better performance in identifying patients at risk of short-term COVID-19 reinfection while minimizing false positives (AUPRC = 0.54, 95% CI 0.46–0.61), followed by XGBoost (AUPRC = 0.49, 95% CI: 0.43–0.55) and glmnet models (AUPRC = 0.47, 95% CI 0.43–0.52).

Feature importance based on SHAP value

The SHAP importance score and absolute value of each predictor variable were also computed and shown in Fig. 2. From Fig. 2A, we discovered that the GAD-7 score was one of the most important predictor followed by patient self-report outcome (previous infection), the FACIT-F score, patient self-report outcome (current), the patients with RA, age, the PSQI score, csDMARDs taper during initial acute infection, the number of symptoms during initial acute infection, the PHQ-9 score, the patients with connective tissue diseases (CTD), treatment with csDMARDs before initial infection, glucocorticoid taper during initial acute infection and being treated with iguratimod before initial infection. From Fig. 2B, the GAD-7 score, patient self-report outcome (previous infection), csDMARDs taper during initial acute infection, the number of symptoms during initial acute infection, glucocorticoid taper during initial acute infection, and being treated with iguratimod before initial infection showed positive SHAP value for the high value (distributed on the right side of 0 baseline). The FACIT-F score, patient self-report outcome (current), the patients with RA, and the patients with CTD showed negative SHAP value for the high value (distributed on the left side of 0 baseline). The specific key SHAP values for the important predictors can be referred to the Supplementary Table 3.

Fixed effect logistic regression based on feature selection

The fixed effect logistic regression results for screening risk variables associated with short-term COVID-19 reinfection in patients with rheumatic diseases were presented in Table 3. In the overall model, the following were significantly associated with higher likelihood of short-term COVID-19 reinfection: having glucocorticoid taper (OR = 2.61, 95% CI 1.38–4.92, e-value = 2.61), having csDMARDs taper (OR = 2.97, 95% CI 1.90–4.64, e-value = 2.84), the number of symptoms (OR = 1.24, 95% CI 1.08–1.42, e-value = 1.47), GAD-7 scores (OR = 1.07,

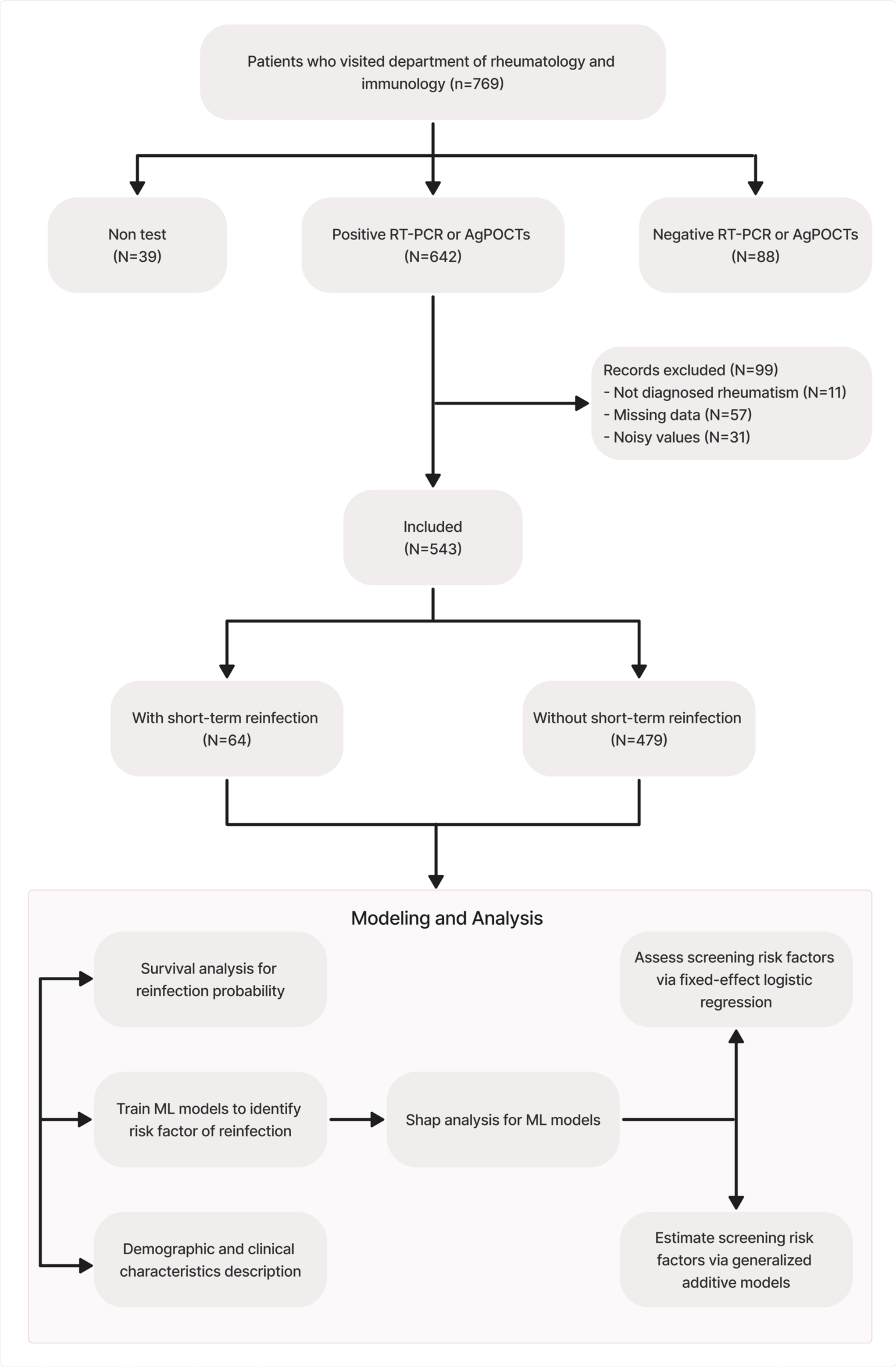
Patients who visited department of rheumatology and immunology (n=769)
Non test (N=39)
Positive RT-PCR or AgPOCTs (N=642)
Negative RT-PCR or AgPOCTs (N=88)
Records excluded (N=99)
- Not diagnosed rheumatism (N=11)
- Missing data (N=57)
- Noisy values (N=31)
Included (N=543)
With short-term reinfection (N=64)
Without short-term reinfection (N=479)
Modeling and Analysis
Survival analysis for reinfection probability
Train ML models to identify risk factor of reinfection
Shap analysis for ML models
Demographic and clinical characteristics description
Assess screening risk factors via fixed-effect logistic regression
Estimate screening risk factors via generalized additive models

Fig. 1 Flowchart for the overall study design and modeling analysis steps

Table 2 Machine learning models performance

Models	AUC (95% CI)	AUPRC (95% CI)	AUPRC baseline	Gmean
LightGBM	0.88 (0.83–0.92)	0.42 (0.36–0.48)	0.12	0.48
XGBoost	0.91 (0.87–0.95)	0.49 (0.43–0.55)	0.12	0.76
Random Forest	0.89 (0.85–0.93)	0.54 (0.46–0.61)	0.12	0.39
Glmnet	0.86 (0.78–0.94)	0.47 (0.43–0.52)	0.12	0.54

AUC Area Under the Curve, *AUPRC* Area Under the Precision-Recall Curve. Gmean means the sqrt (sensitivity * specificity). 95% CI shows the uncertainty for AUC and AUPRC metrics. AUPRC baseline means the value that a random classifier would achieve (random guessing)

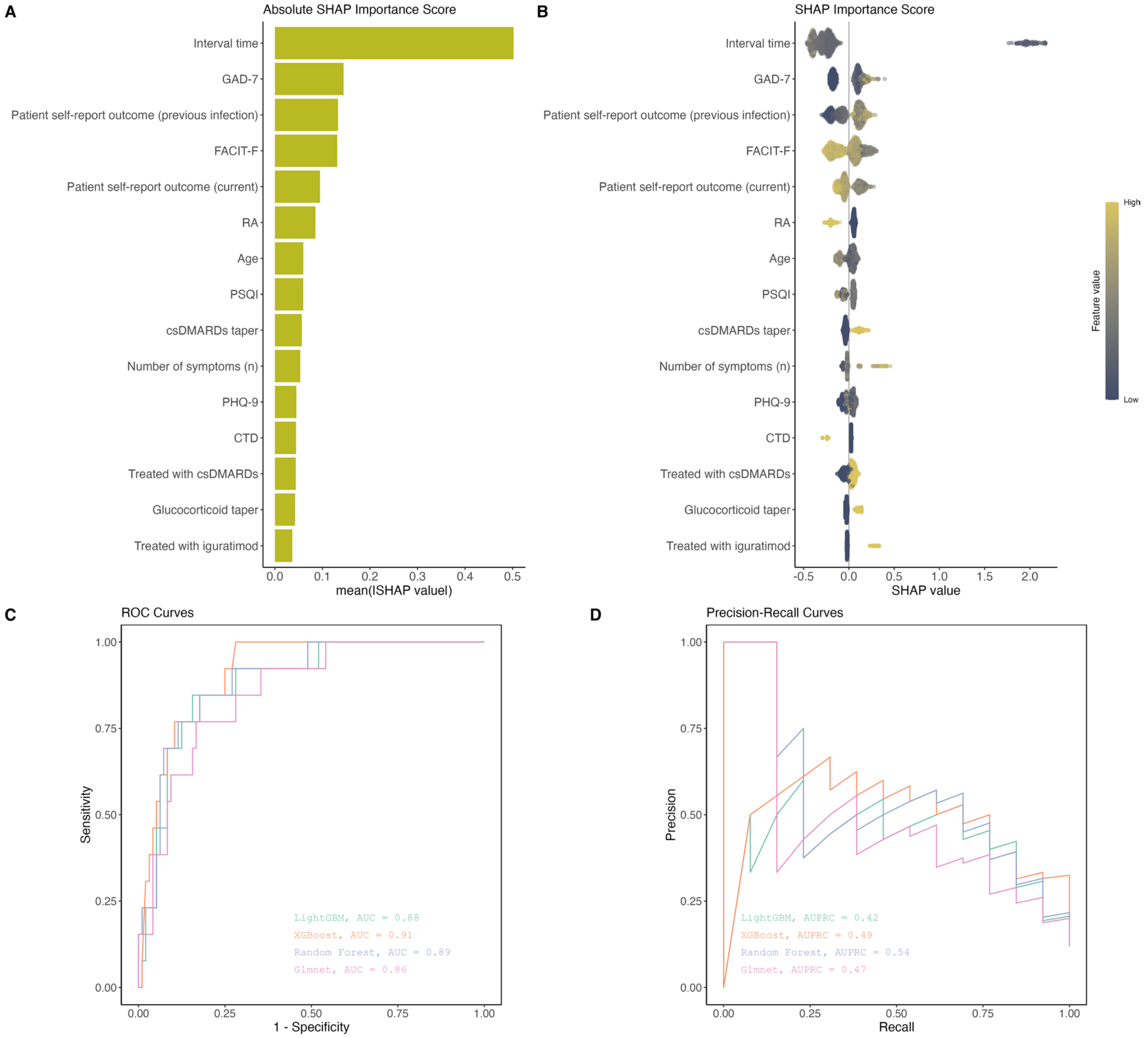

Fig. 2 SHAP importance and model performance plots. **A** Represents the mean absolute SHAP values for each predictive variable. **B** Represents the distribution of SHAP values for predictive variables. X-axis showed the direction of the importance score (negative or positive). Colors change from low values (dark blue) to high values (yellow). **C** ROC curves, represents the model performance on the classification. **D** Precision-recall curves

Table 3 Regression results for screening risk variables

	Odds ratio (95% CI)	P value	P adjust	E-value
Demographics				
Sex				
Male	[Reference]			
Female	0.69 (0.33–1.44)	0.323	0.420	1.69
Age	0.98 (0.95–1.01)	0.231	0.334	1.11
csDMARDs				
Treated with csDMARDs				
N	[Reference]			
Y	0.77 (0.29–2.09)	0.611	0.611	1.53
Treated with Iguratimod				
N	[Reference]			
Y	1.46 (0.52–4.02)	0.469	0.508	1.71
Drug adjustment				
Glucocorticoid taper				
N	[Reference]			
Y	2.61 (1.38–4.92)	0.003	0.010	2.61
csDMARDs taper				
N	[Reference]			
Y	2.97 (1.90–4.64)	< 0.001	< 0.001	2.84
Adverse reactions				
Number of symptoms	1.24 (1.08–1.42)	0.002	0.008	1.47
Psychological health surveys				
FACIT-F	0.95 (0.93–0.96)	< 0.001	< 0.001	1.19
PHQ-9	1.06 (0.99–1.12)	0.074	0.120	1.20
GAD-7	1.07 (1.02–1.13)	0.004	0.010	1.23
PSQI	1.06 (1.00–1.13)	0.060	0.112	1.21
Health status score				
Patient self-report outcome (previous infection)	1.06 (0.92–1.22)	0.447	0.528	1.20
Patient self-report outcome (current)	0.87 (0.76–0.99)	0.041	0.089	1.35

P-adjusted values are determined using the Benjamini-Hochberg (BH) correction method. The E-value measures the minimal magnitude of an association necessary, similar in scale to the observed effect size, for a potential confounding variable to explain the observed association

95% CI 1.02–1.13, e-value = 1.23). The following were significantly associated with lower likelihood of short-term COVID-19 reinfection: FACIT-F scores (OR = 0.95, 95% CI 0.93–0.96, e-value = 1.19).

Survival analysis and non-linear pattern analysis

Through Kaplan–Meier survival analysis, we found that the probability of COVID-19 short-term reinfection was similar in rheumatic patients with different sex (Fig. 3). The generalized additive models and smooth curve fitting were also conducted to clarify the potential non-linear relationship between psychological status, quality of sleep, patient self-report outcome and the short-term reinfection probability of patients with rheumatic diseases (Fig. 3). A higher FACIT-F score indicated a lower fatigue level. On the contrary, the higher the scores of GAD-7 scale, PHQ-9 scale, PSQI scale and EQ-5D-VAS scale indicated a higher anxiety level, a higher depression level, a worse the sleep quality and a higher disease activity. We found that patients' risk for short-term reinfection gradually decreased with increasing FACIT-F scores (P = 0.024) and EQ-5D VAS scores (current) (P = 0.044), while it gradually increased with increasing GAD-7 scores (P = 0.003). Although there was no statistical difference in PHQ-9, PSQI and EQ-5D VAS scores (first infection), it showed a trend that depression, bad sleep quality and a higher disease activity at previous infection contributed to higher risk for short-term reinfection.

Discussion

Given the threat that short-term reinfection of COVID-19 poses to disease control and clinical outcomes in people with underlying diseases and immune disorders, it is necessary to predict the risk of short-term reinfection of COVID-19 in the context of medical care for rheumatic diseases. In a few studies, machine learning has been

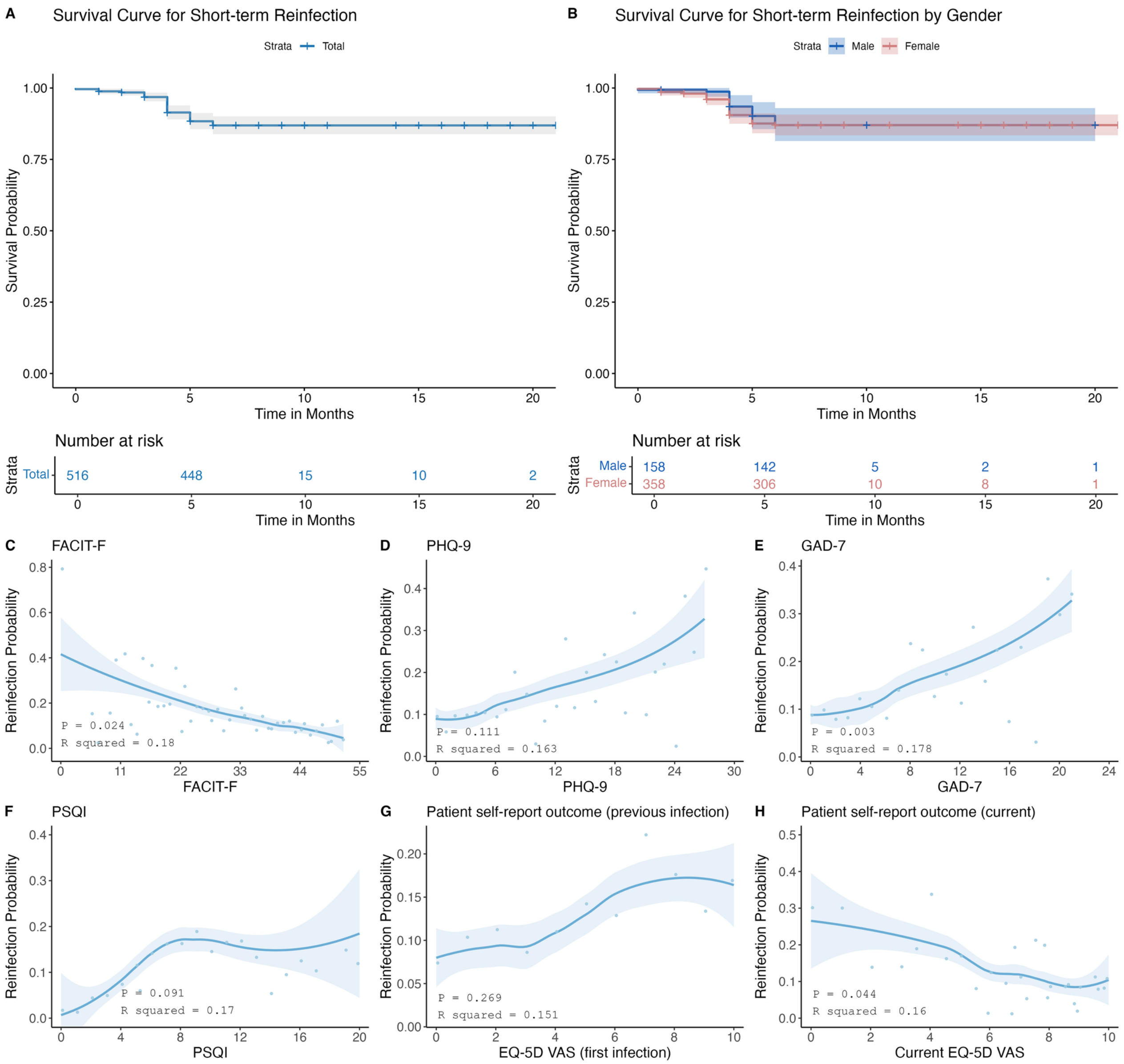

Fig. 3 Survival curves and GAM trajectories plots. **A** and **B** represent the probability curves for short-term reinfection for the total dataset and sex subgroups, respectively. **C**–**H** represent estimated probability of reinfection based on changes of FACIT-F, PHQ-9, GAD-7, PSQI, patient self-report outcome (previous infection), patient self-report outcome (current), respectively. Lightblue part means the 95% CI for each blue trajectories line

applied to predict COVID-19 reinfection based on biological indicators [32] or radiomic data [33] and seemed to show potential. In addition, machine learning has also been applied to predict overall survival of reinfected patients [34] and screened characteristics of hospital admissions due to COVID-19 reinfection [35]. Factors associated with COVID-19 reinfection have also been preliminarily explored [36–38]. However, this problem in patients with rheumatic diseases has not been effectively dealt with. Some studies have shown that adverse psychological health problems not only delay clinical recovery from COVID-19 [39], but also have a negative impact on the prognosis of COVID-19-related cardiovascular disease [40]. Nevertheless, few studies included psychological scale scores to assist to predicting COVID-19 reinfection. In this study, we focused on rheumatic disease condition, rheumatic disease control, COVID-19 related symptoms, adverse reactions to COVID-19 vaccine, influenza vaccination, drug dose adjustment during acute COVID-19 infection, and psychological factors.

We compared the performance of four machine learning algorithms in predicting the short-term reinfection risk of COVID-19 and identified the key variables in the data that contributed to the short-term reinfection risk of COVID-19. It was finally found that the XGBoost algorithm produced better performance, and it could make informed predictions (AUC = 0.91; 95% CI 0.87–0.95). Substantial factors of short-term reinfection included glucocorticoid taper, csDMARDs taper, a large number of symptoms, high GAD-7 scores, and low FACIT-F scores.

The prediction of COVID-19 short-term reinfection was necessary for patients with rheumatic diseases, but there was little literature on it. We found that patients with younger age were more likely to suffer short-term reinfection. The reason may be that people at this age were busier and more stressed at work than older people. They were easier to be immunocompromised due to the lack of sleep, anxiety, and other similar conditions. The opinions of Alexander Lawandi et al. [41], Nicole Bechmann et al. [42], Emily N. Kowalski et al. [43] and Chen Yi-Hsuan et al. [44] were consistent with ours. However, according to a recent research, individuals over 60 years of age appeared to be more likely to be reinfected when exposed to a new variant [45]. The reason may be that that study only reviewed the results of 30 articles and could not fully reflect the true situation. Overall, patients with younger age were at higher risk to suffer short-term reinfection. In addition, similar to previous studies [39, 46–54], we found that patients with poor psychological health, sleep quality, and a larger number of symptoms tended to suffer short-term reinfection. The potential reason may be that poor psychological health and sleep quality led to the weakened immune system and disrupted the hormone level which led to the decline of human resistance to the virus. Fatigue status are associated with an increased risk of COVID-19 recurrence [46]. Besides, the study of Xing Wang et al. also included urgent attention to the depression and insomnia of re-positive patients [47]. Maojun Li et al. identified depressive status as independent risk factors for re-positivity [39]. Although these studies have all shown that poor psychological health and sleep quality were associated with developing COVID-19, none have addressed the role of psychological factors and sleep quality in predicting short-term reinfection.

Results of a multicenter study showed that the number of symptoms during acute infection was associated with long COVID-19 [48]. Many subsequent research results also provided support for this view [49, 50]. Similarly, the study of Schmidbauer Lena et al. revealed the promoting effect of the number of symptoms and persistent symptoms during acute infection on post-COVID-19 fatigue [55]. Meanwhile, with the increase of the number of symptoms, the risk of combining with depression may also be increased [52]. Both Bilgin Aylin et al. [53] and Adar Sevda et al. [54] evaluated pain using VAS scale and found that higher degree of pain appears to be also associated with chronic fatigue and anxiety levels, but significant changes have not been observed in disease activity following COVID-19 reinfection in patients with RA [55]. They promoted each other. However, the relationship between the number of symptoms during acute infection, patient self-report outcome, and the risk of reinfection has not been effectively explored. On balance, similar to current evidence, younger age at initial acute infection, anxiety, fatigue, depression, poor quality of sleep, and a higher number of COVID-19 related symptoms during initial acute infection may be associated with a higher risk of short-term reinfection. Further exploration is also needed on the relationship between patient self-report outcome and reinfection.

We also discovered that patients with RA and CTD may be associated with a lower risk of short-term reinfection, while glucocorticoid taper and csDMARDs taper may be associated with a higher risk of short-term reinfection. However, this view was not supported by much evidence. Similar to our view, Elena Beyzarov et al. found that in cases involving potential COVID-19 reinfection, the proportion of immunosuppressants or immunomodulators reported was relatively low [56]. This result appears to be at odds with the common view that immunosuppression is a risk factor which increases susceptibility to COVID-19 infection and thus promotes reinfection [34, 57–60]. A review identified prednisolone as a potential cause of COVID-19 recurrence [61] while the other study reported that there was no statistical difference in the use of corticosteroid therapy in patients with or without recurrence [62]. Our views did not coincide with theirs. Mehran Pournazari et al. through the analysis of the clinical features of infection found that infected patients have the highest percentage of RA reports [63]. In summary, the potential role of specific rheumatic disease entities, glucocorticoid and immunosuppressants remain controversial and need to be evaluated in more studies. To exploration of the implications of the SHAP analysis deeply, we comparised with the existing literature and more details were shown in Supplementary Table 2.

Comprehensive treatment measures should be taken for the management of patients with rheumatic diseases after infection, such as early identification of high-risk patients, focusing on clinical care, mental health, and improvement of quality of sleep. In the actual clinical work, medical staff should advise patients to keep a regular schedule and adequate sleep, and avoid fatigue. After infection, the adjustment of glucocorticoid and csDMARDs should be careful. When the pain is severe, appropriate use of drugs is necessary to relieve pain. Psychological care is also necessary to relieve anxiety and depression by listening, understanding and encouragement.

Limitations

This article has several limitations. Firstly, participants were recruited from a hospital in Guangzhou, which led to potential sampling bias. Secondly, we can not conclude that the association between these factors is real causality because the study is cross-sectional. Thirdly, the application of a self-rating scale to assess psychological status is the limitation of this study. Fourthly, additional external validation methods have not been used to verify our models' performance due to the complex variable sources. Finally, the generalizability of the models among different populations or healthcare settings is still unknown and can be explored in the future.

Conclusion

We developed a robust machine learning model through the XGBoost algorithm to predict the risk of short-term COVID-19 reinfection in patients with rheumatic diseases. SHAP analysis highlighted the importance of clinical and psychological factors. Factors such as anxiety, fatigue, depression, poor sleep quality, high disease activity during initial infection, and the use of glucocorticoid taper were significant predictors. For the patient population with above factors, more attention should be paid to preventing COVID-19 short-term reinfection. Machine learning may assist in further screening potential risk factors that influence the likelihood of short-term COVID-19 reinfection among patients with rheumatic diseases. In the future, the performance accuracy of our model and its generalizability would be enhanced if the larger and multicenter dataset containing inflammatory markers data were integrated into the model.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-024-05982-2.

Supplementary material 1.

Acknowledgements
Not applicable.

Author contributions

SX, YL, and YJ developed the study design. SX, YL, and YJ led data collection and interpretation. SX,YL drafted and revised the manuscript. SX, YL, XZ, XW, SD, and YJ reviewed and revised the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by National Natural Science Foundation of China Youth Fund [82201998], 2024 Basic and Applied Basic Research (Outstanding Doctoral Program) [2024A04J9909], Guangdong Provincial Basic and Applied Basic Research Fund Project [2024A1515012734] and Third Affiliated Hospital of Sun Yat-Sen University Second Major Talent Development Program.

Availability of data and materials

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The participants provided their written informed consent to participate in this study. Aligned with the Helsinki Declaration, ethical approval of the Institutional Review Board (IRB) for the current study was obtained from the third affiliated hospital of Sun Yat-sen university ethical committee (Number: II2023-090–02).

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests to disclose.

Author details
[1] Department of Rheumatology and Immunology, Third Affiliated Hospital of Sun Yat-Sen University, 600 Tianhe Road, Tianhe District, Guangzhou, China. [2] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [3] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

Received: 18 July 2024 Accepted: 13 December 2024
Published online: 24 December 2024

References

1. Van Elslande J, Vermeersch P, Vandervoort K, et al. Symptomatic Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) reinfection by a phylogenetically distinct strain. Clin Infect Dis. 2021;73:354.
2. Edridge AWD, Kaczorowska J, Hoste ACR, et al. Seasonal coronavirus protective immunity is short-lasting. Nat Med. 2020;26:1691.
3. Zhang M, Cao L, Zhang L, et al. SARS-CoV-2 reinfection with Omicron variant in Shaanxi Province, China: December 2022 to February 2023. BMC Public Health. 2024;24:496.
4. Cai C, Li Y, Hu T, et al. The associated factors of SARS-CoV-2 reinfection by omicron variant - Guangdong Province, China, December 2022 to January 2023. China CDC Wkly. 2023;5:391–6.
5. Liu D, Chen B, Liao X, et al. Specific persistent symptoms of COVID-19 and associations with reinfection: a community-based survey study in southern China. Front Public Health. 2024. https://doi.org/10.3389/fpubh.2024.1452233.
6. Reinfection. https://www.cdc.gov/coronavirus/2019-ncov/your-health/reinfection.html.
7. Romero-Ibarguengoitia ME, Rodríguez-Torres JF, Garza-Silva A, Rivera-Cavazos A, Morales-Rodriguez DP, Hurtado-Cabrera M, et al. Association of vaccine status, reinfections, and risk factors with Long COVID syndrome. Sci Rep. 2024;14:2817.
8. Strand V, Simon LS, Meara AS, Touma Z. Measurement properties of selected patient-reported outcome measures for use in randomised controlled trials in patients with systemic lupus erythematosus: a systematic review. Lupus Sci Med. 2020;7:e000373.
9. Cella D, Lenderking WR, Chongpinitchai P, Bushmakin AG, Dina O, Wang L, et al. Functional assessment of chronic illness therapy-fatigue is a reliable and valid measure in patients with active ankylosing spondylitis. J Patient Rep Outcomes. 2022;6:100.
10. Wang S-Y, Zang X-Y, Liu J-D, Gao M, Cheng M, Zhao Y. Psychometric properties of the functional assessment of chronic illness therapy-fatigue

(FACIT-Fatigue) in Chinese patients receiving maintenance dialysis. J Pain Symptom Manage. 2015;49:135–43.

11. Wang W, Bian Q, Zhao Y, Li X, Wang W, Du J, et al. Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry. 2014;36:539–44.
12. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166:1092–7.
13. Chong AML, Cheung C. Factor structure of a Cantonese-version Pittsburgh sleep quality index. Sleep Biol Rhythms. 2012;10:118–25.
14. Zhou T, Guan H, Yao J, Xiong X, Ma A. The quality of life in Chinese population with chronic non-communicable diseases according to EQ-5D-3L: a systematic review. Qual Life Res. 2018;27:2799–814.
15. Yao Q, Liu C, Zhang Y, Xu L. Population norms for the EQ-5D-3L in China derived from the 2013 National Health Services Survey. J Glob Health. 2021. https://doi.org/10.7189/jogh.11.08001.
16. Sun S, Chen J, Kind P, Xu L, Zhang Y, Burström K. Experience-based VAS values for EQ-5D-3L health states in a national general population health survey in China. Qual Life Res. 2015;24:693–703.
17. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Neural Inf Process Syst. 2017.
18. 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. ACM: San Francisco California USA, 2016; 785–794.
19. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2:18–22.
20. Yuan G-X, Ho C-H, Lin C-J. An improved GLMNET for l1-regularized logistic regression. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: San Diego California USA, 2011; 33–41.
21 Berrar D. Cross-validation. Amsterdam: Elsevier; 2019.
22. Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Appl. 2022;44:875–86.
23. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr. 2008;17:145–51.
24. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015;68:855–9.
25 Tang Y, Zhang Y-Q, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern Part B. 2008;39:281–8.
26. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. pp. 4768–77.
27. Allison PD. Fixed effects regression models. Thousand Oaks: SAGE publications; 2009.
28. Ferreira J, Zwinderman A. On the Benjamini-Hochberg method. Ann Stat. 2006. https://doi.org/10.1214/009053606000000425.
29. Blum MR, Tan YJ, Ioannidis JP. Use of E-values for addressing confounding in observational studies—an empirical assessment of the literature. Int J Epidemiol. 2020;49:1482–94.
30. Hastie TJ. Generalized additive models. In: Chambers JM, Hastie TJ, editors. Statistical models in S. Oxfordshire: Routledge; 2017. p. 249–307.
31. Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. J Am Stat Assoc. 1988;83:414–25.
32. Chen J, Luo D, Sun C, Sun X, Dai C, Hu X, et al. Predicting COVID-19 re-positive cases in malnourished older adults: a clinical model development and validation. CIA. 2024;19:421–37.
33. Wang X-H, Xu X, Ao Z, Duan J, Han X, Tang X, et al. Elaboration of a radiomics strategy for the prediction of the re-positive cases in the discharged patients with COVID-19. Front Med. 2021;8:730441.
34. Ebrahimi V, Sharifi M, Mousavi-Roknabadi RS, et al. Predictive determinants of overall survival among re-infected COVID-19 patients using the elastic-net regularized Cox proportional hazards model: a machine-learning algorithm. BMC Public Health. 2022;22:10.
35. Afrash MR, Kazemi-Arpanahi H, Shanbehzadeh M, et al. Predicting hospital readmission risk in patients with COVID-19: A machine learning approach. Inform Med Unlocked. 2022;30:100908.
36. Li X, Yin D, Yang Y, Bi C, Wang Z, Ma G, et al. Eosinophil: a nonnegligible predictor in COVID-19 re-positive patients. Front Immunol. 2021;12:690653.
37. Chen LZ, Lin ZH, Chen J, Liu SS, Shi T, Xin YN. Can elevated concentrations of ALT and AST predict the risk of 'recurrence'of COVID-19? Epidemiol Infect. 2020;148:e218.
38. Zheng Y, Wang J, Ding X, Chen S, Li J, Shen B. The correlation between triglyceride-glucose index and SARS-CoV-2 RNA re-positive in discharged COVID-19 patients. IDR. 2022;15:3815–28.
39. Li M, Peng H, Duan G, Wang J, Yu Z, Zhang Z, et al. Older age and depressive state are risk factors for re-positivity with SARS-CoV-2 omicron variant. Front Public Health. 2022;10:1014470.
40 Gonjilashvili A, Tatishvili S. The interplay between Sars-Cov-2 infection related cardiovascular diseases and depression. Common mechanisms, shared symptoms. Am Heart J Plus. 2024;38:100364.
41. Lawandi A, Warner S, Sun J, Demirkale CY, Danner RL, Klompas M, et al. Suspected Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2) reinfections: incidence, predictors, and healthcare use among patients at 238 US healthcare facilities, 1 June 2020 to 28 February 2021. Clin Infect Dis. 2022;74:1489–92.
42. Bechmann N, Barthel A, Schedl A, Herzig S, Varga Z, Gebhard C, et al. Sexual dimorphism in COVID-19: potential clinical and public health implications. Lancet Diabetes Endocrinol. 2022;10:221–30.
43. Kowalski EN, Wang X, Patel NJ, Kawano Y, Cook CE, Vanni KM *et al.* Risk factors and outcomes for repeat COVID-19 infection among patients with systemic autoimmune rheumatic diseases: a case-control study. In: Seminars in Arthritis and Rheumatism. Elsevier. 2023; 152286.
44. Chen YH, Lee CY, Cheng HY, et al. Risk factors and mortality of SARS-CoV-2 reinfection during the Omicron era in Taiwan: a nationwide population-based cohort study. J Microbiol Immunol. 2024;57:30.
45. Gómez-Gonzales W, Chihuantito-Abal LA, Gamarra-Bustillos C, Morón-Valenzuela J, Zavaleta-Oliver J, Gomez-Livias M, et al. Risk factors contributing to reinfection by SARS-CoV-2: a systematic review. Adv Respir Med. 2023;91:560–70.
46. Hoang T. Systematic review and meta-analysis of factors associated with re-positive viral RNA after recovery from COVID-19. J Med Virol. 2021;93:2234–42.
47. Wang X, Fan Q, Li Y, Xiao J, Huang Y, Guo T, et al. The changes in psychological symptoms of COVID-19 patients after "re-positive." Front Psych. 2022;13:1010004.
48. Fernández-de-Las-Peñas C, Pellicer-Valero OJ, Navarro-Pardo E, Palacios-Ceña D, Florencio LL, Guijarro C, et al. Symptoms experienced at the acute phase of SARS-CoV-2 infection as risk factor of long-term post-COVID symptoms: the LONG-COVID-EXP-CM multicenter study. Int J Infect Dis. 2022;116:241–4.
49. Kandemir H, Bülbül GA, Kirtiş E, Güney S, Sanhal CY, Mendilcioğlu İİ. Evaluation of long-COVID symptoms in women infected with SARS-CoV -2 during pregnancy. Intl J Gynecol Obste. 2024;164:148–56.
50. Ko ACS, Candellier A, Mercier M, Joseph C, Schmit J-L, Lanoix J-P, et al. Number of initial symptoms is more related to long COVID-19 than acute severity of infection: a prospective cohort of hospitalized patients. Int J Infect Dis. 2022;118:220–3.
51. Schmidbauer L, Kirchberger I, Goßlau Y, Warm TD, Hyhlik-Dürr A, Linseisen J, et al. The association between the number of symptoms and the severity of Post-COVID-Fatigue after SARS-CoV-2 infection treated in an outpatient setting. J Neurol. 2023;270:3294–302.
52. Shah A, Bhattad D. Immediate and short-term prevalence of depression in covid-19 patients and its correlation with continued symptoms experience. Indian J Psychiatry. 2022;64:301–6.
53. Bilgin A, Kesik G, Özdemir L. Biopsychosocial factors predicting pain among individuals experiencing the novel coronavirus disease (COVID-19). Pain Manag Nurs. 2022;23:79–86.
54. Adar S, Konya PŞ, Akçin Aİ, Dündar Ü, Demirtürk N. Evaluation and follow-up of pain, fatigue, and quality of life in COVID-19 patients. Osong Public Health Res Perspect. 2023;14:40.
55. Kim YE, Ahn SM, Oh JS, et al. Prevalence and risk factors of COVID-19 reinfection in patients with rheumatoid arthritis: a retrospective observational study. Yonsei Med J. 2024;65:645–50.
56. Beyzarov E, Chen Y, Caubel P. Reporting of COVID-19 reinfection and potential role of immunosuppressant/immunomodulating agents: a

cross-sectional observational analysis based on a spontaneous reporting database. Clin Drug Investig. 2022;42:807–12.

57. SotoodehGhorbani S, Taherpour N, Bayat S, Ghajari H, Mohseni P, HashemiNazari SS. Epidemiologic characteristics of cases with reinfection, recurrence, and hospital readmission due to COVID-19: a systematic review and meta-analysis. J Med Virol. 2022;94:44–53.
58 Zheng YQ, Li HJ, Chen L, et al. Immunogenicity of inactivated COVID-19 vaccine in patients with autoimmune inflammatory rheumatic diseases. Sci Rep. 2022. https://doi.org/10.1038/s41598-022-22839-0.
59. Yousefghahari B, Navari S, Sadeghi M, et al. Risk of COVID-19 infection in patients with rheumatic disease taking disease-modifying anti-rheumatic drugs. Clin Rheumatol. 2021;40:4309–15.
60. Hunsinger DHP, KuttiSridharan DG, Rokkam DVRP, et al. COVID-19 reinfection in an immunosuppressed patient without an antibody response. Am J Med Sci. 2021;362:103.
61. Piri SM, Edalatfar M, Shool S, Jalalian MN, Tavakolpour S. A systematic review on the recurrence of SARS-CoV-2 virus: frequency, risk factors, and possible explanations. Infect Dis. 2021;53:315–24.
62. Zou Y, Wang B-R, Sun L, Xu S, Kong Y-G, Shen L-J, et al. The issue of recurrently positive patients who recovered from COVID-19 according to the current discharge criteria: investigation of patients from multiple medical institutions in Wuhan, China. J Infect Dis. 2020;222:1784–8.
63. About Reinfection. COVID-19. CDC; 2024.

Publisher's Note