scientific reports

Check for updates

OPEN

Development and internal validation of a prediction model for rheumatoid arthritis: a case-control study

Ling Tu[1,4], Fuling Wei[1,4], Yuqing Song[2], Haitao Huang[1], Ligang Qing[3], Xi Luo[3], Ying Liu[1] & Hong Chen[1]✉

This study measured sociodemographic characteristics, dietary habits, lifestyle habits, genetics, and other factors that may contribute to the development of Rheumatoid Arthritis (RA). Independent risk factors for RA were identified by logistic regression analysis, and a prediction model was constructed. The area under the receiver operating characteristic curve (AUC) was used to evaluate the prediction accuracy of the model, and the calibration of the model was evaluated by the Hosmer–Lemeshow test. A total of 432 participants, comprising 216 healthy individuals and 216 patients diagnosed with RA at two hospitals in Sichuan, China, from March 2022 to January 2023 were included in this study. Logistic regression analysis revealed that occupation type, place of residence, history of mumps, dietary combination, sweet, damp dwelling, fish, vaccine history, and rs805297 were significantly associated with the pathogenesis of RA. The model constructed in this study showed good prediction, with an AUC of 0.912. The Youden index was 0.699, the sensitivity was 0.847 and the specificity was 0.852. The Hosmer–Lemeshow test results ($\chi^2 = 8.441$, $P = 0.392$) indicated that the model had good diagnostic value. The internal validation AUC was 0.942. We propose a new promising model for identifying individuals at risk of developing RA.

Keywords Rheumatoid arthritis, Risk prediction model, Risk factors, Risk assessment, Case-control study

Rheumatoid arthritis (RA) is a chronic, systemic, highly disabling autoimmune disease. The main clinical manifestation is aggressive joint inflammation[1]. An epidemiological survey showed that the global incidence of RA is approximately 0.5%~1%[2], and the incidence of RA in China is 0.42%, with a total number of patients of approximately 5 million[3]. A study has shown that the mortality rate of RA patients is 1.6 times greater than that of the general population[4]. RA has the characteristics of a prolonged disease course and repeated disease, which can easily lead to a reduction in patients' daily living function, quality of life, and social participation. It can also cause serious economic burdens to patients, families, and society[5–7]. Clarifying the risk factors for developing RA can help clinical staff identify people at risk for RA at an early stage. Individualized prevention programs and early interventions based on risk levels are valuable for reducing the incidence of this disease.

The pathogenesis of RA is affected by many factors. Currently, it is believed that the combination of genetic risk factors and environmental risk factors is the main cause of RA pathogenesis[8,9]. Several studies at home and abroad have shown that genetic factors are risk factors for developing RA and account for 50–65% of susceptibility to RA[5,10–12]. Epidemiologic investigations have shown that risk factors affecting the pathogenesis of RA include environmental risk factors, such as smoking[5,12,13], silica dust exposure[14], humid living conditions, and air pollution[15,16]. In addition, dietary factors such as high sodium intake[17], coffee intake[18], and meat intake[19]are also risk factors for the development of RA. In addition, several comorbidities, such as diabetes mellitus[20]and periodontitis[21], are risk factors for developing RA. Other risk factors, such as education level, alcohol consumption, occupation[22], menopause[23], and ultraviolet light[8,24], also have an impact on the development of RA.

[1]Department of Nursing, West China School of Nursing, West China Hospital, Sichuan University, Chengdu 610064, China. [2]School of Nursing, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China. [3]Shi Yang Community Health Service Center, Chengdu 610041, China. [4]Ling Tu and Fuling Wei contributed equally. ✉email: 1366109878@qq.com

However, due to the complex and diverse pathogenesis of RA, the influence of several factors on RA pathogenesis is unclear[8,25]. To accurately screen high-risk groups for disease prevention, it is necessary to establish a prediction model that can predict individual RA risk.

At present, risk prediction models have been developed in other countries. The RA genetic risk scoring system included only genetic loci. The RA risk prediction model based on asymptomatic individuals incorporates both genetic and environmental factors. The RA risk prediction model based on arthralgia incorporates serologic markers and clinical symptoms. Foreign RA incidence risk prediction models are mostly for Caucasian populations and include different risk factors. The area under the receiver operating characteristic curve (AUC) of the prediction models ranged from 0.56 to 0.86, with varying diagnostic sensitivities. Those with the largest AUC values were included in studies of seropositive arthralgia patients, targeting a specific population group[24,26–33]. On the other hand, most of the studies on the risk factors for RA in China are cross-sectional[34], the level of evidence is low, and there is no research on risk prediction models for RA. There are only predictive models for complications such as RA combined with interstitial lung disease[35] and combined infection.

The diagnostic sensitivity of these models varies, and their effectiveness in different cohorts needs to be validated. The detection of several inclusion indicators is complex, resulting in less generalizable models. Furthermore, due to differences in ethnicity, culture, geography, and living environment, it is necessary to construct a risk prediction model for RA patients in China[29,31,36].

Methods

Study population

We recruited 216 RA patients from the Department of Rheumatology and Immunology of two tertiary hospitals in Sichuan Province, China, as the case group from March 2022 to January 2023. We also recruited 216 healthy people from a community health service center in Sichuan Province, China, as the control group. In this study, the case group and the control group were matched 1:1 according to sex and age (± 3 years).

Inclusion and exclusion criteria

The inclusion criteria for RA patients were as follows: ① According to the 2010 ACR/EULAR classification of RA[37], patients who were first diagnosed with RA by clinicians were included. Specifically, (a) clinical synovitis in at least one joint; (b) synovitis that cannot be explained by other diseases; and (c) no typical bone erosion on X-ray. If the above three conditions are met, a 4-item score is given, and RA is diagnosed when the total score is ≥ 6. ② Patients ≥ 18 years of age; ③ Patients provided informed consent to participate in the study voluntarily.

The exclusion criteria for RA patients were as follows: ① comorbidity with other rheumatologic diseases, such as ankylosing spondylitis, systemic lupus erythematosus, psoriasis, mixed connective tissue disease, dry syndrome, dermatomyositis, polymyositis, rheumatic polymyalgia, ANCA-related vasculitis, polyarteritis nodosa, leukoaraiosis, or recurrent polychondritis; and ② severe mental illness or cognitive impairment who were unable to cooperate with the study.

The inclusion criteria for the control group were as follows: ① ≥18 years of age; and ② provided informed consent and voluntary participation in the study.

The exclusion criteria for the control group were as follows: ① incomplete questionnaire data; ② diagnosed with RA; and ③ suffering from joint diseases or other connective tissue diseases.

Variables

The investigators collated and analyzed the results of studies on RA risk factors and consulted with rheumatology clinicians and nursing professionals to design their own unified questionnaire. The questionnaire included the following:

Sociodemographic information: age, sex, ethnicity, marital status, education, place of residence, height, weight, type of occupation, working status, per capita income, etc.

Disease history: first-degree relatives, second-degree relatives, allergies, mumps, tuberculosis, chronic disease, periodontal disease, surgery, trauma, blood transfusion, etc.

Dietary status: dietary patterns, regularity of meals, type of taste, intake of dairy products, fresh vegetables, fruits, preserved foods, fish, etc.

Lifestyle: smoking status; frequency of smoking; alcohol consumption; frequency of alcohol consumption; exposure to secondhand smoke, tea, or coffee; physical exercise; frequency of exercise; mode of exercise; etc.

The environmental factors included whether the working environment was humid, the living floor, the living environment was humid, and whether the individual was exposed to dust.

The genetic factors included MHC-related genes and non-MHC-related genes (rs10499494, rs11203366, rs11203367, rs1544410, rs1748033, rs1800796, rs2004640, rs2240340, rs2488457, rs3087243, rs4810485, rs5029937, rs5744280, rs6920220, rs729302, rs7528684, rs7975232, rs805296, rs805297, rs874881, rs9272219, HLA0405, HLA0901 and HLA0401).

For every patient, general demographic information, history of disease, dietary status, lifestyle, environmental factors, genetic factors, and so on were collected. Three milliliters of peripheral blood were collected in EDTA anticoagulant tubes by senior nurses. The whole process strictly followed the Guidelines for Venous Blood Specimen Collection. Commercial DNA extraction kits are used to extract DNA from blood samples. All DNA samples were tested for OD values using a NanoDrop2000 instrument, and 1.25% agarose gel electrophoresis was performed at the same time. Finally, a DNA quality assessment was performed to determine whether the sample met the Massarray SNP typing DNA quality criteria. All the SNP sites selected were genotyped on the mass spectrometry analysis platform (MALDI-TOF), and TYPER4.0 software was used to obtain all the raw data and related genotyping maps of the test results.

Calculation of sample size

The sample size was calculated with the 1:1 individual matching design recommended by Schlessman:

$$m = \left[z_\alpha /2 + z_\beta \sqrt{p\,(1-p)}\right]^2 /(p-0.5)^2 \quad (1)$$

$$M = m/p_0\,(1-p_1) + p_1\,(1-p_0) \quad (2)$$

$$p = OR/(1+OR) \quad (3)$$

$$p_1 = p_0 \bullet\ OR/\,[1 + p_0\,(\mathrm{OR}-1)] \quad (4)$$

m is the number of pairs with inconsistent results, and *M* is the total number of pairs. p_0 and p_1 represent the estimated exposure rates of the control and case groups in the target population, respectively.

In this study, genetic factors, smoking status, and other exposure factors were selected for sample size calculation. Since a high-fat diet needed the largest sample size, a high-fat diet was ultimately selected as one of the exposure factors for sample size estimation. Under the conditions of $\alpha = 0.05$ (two-sided test) and $\beta = 0.10$, according to the relevant surveys, the proportion of individuals exposed in the control group was $p_0 = 0.217$. In a 10-year dietary survey of 25,630 volunteers, Pattison reported that a high-fat diet increased the risk of arthritis (OR = 2.3)[38]. After the above formula was substituted for PASS, 189 patients were included in the case group, and 189 patients were included in the control group. Considering that 10% of the samples collected may be unqualified or that the main variables are missing, the final sample sizes for the case and control groups are proposed to be 210 cases each[39]. In the previous studies, Lotte Arwen van de Stadt's study had 374 participants, and its AUC was 0.86 (95%CI, 0.80)[33]. Elizabeth W Karlson's study involved 898 participants, its AUC was 0.66[28]. Jeffrey A Sparks's study involved 791 participants, and its AUC was 0.82[32]. It was indicated that the sample size of this study was within the acceptable range.

Statistical analysis

EpiData 3.1 was used to input and collect the data by two people, and IBM SPSS 27.0 and R4.1 were used for the data analysis. Normally distributed quantitative data are described as $\bar{x} \pm s$. Qualitative data are described as the number of patients used and the component ratio. The factors with statistical significance ($P < 0.05$) according to the univariate logistic regression analysis were analyzed via multivariate logistic regression. The AUC was used to evaluate the prediction accuracy of the model, and the calibration of the model was evaluated by the Hosmer–Lemeshow test. Finally, the prediction formula for individual RA risk was determined by the coefficients of the prediction model. Bootstrap validation of the data was performed for internal validation of the model, where each resampling generated a new training set and the unselected samples formed the validation set[40].

Ethics in research

This study was approved by the Ethics Review Committee and conducted in accordance with the Declaration of Helsinki. After review by the Ethics Committee on Biomedical Research, the study officially started with the ethical review number 2021 (175).

Results

Characteristics of the study population

A total of 216 patients with RA who attended two tertiary hospitals in Sichuan Province, China, from March 2022 to January 2023 were selected as the case group, and a total of 216 healthy patients who attended a community health center in Sichuan Province, China, during the same period were selected as the control group. Matching principle: RA patients (case group) and healthy people (control group) were matched 1:1 according to sex and age (± 3 years). The two groups were comparable. Table 1 summarizes the sociodemographic information of the RA patients and healthy individuals.

Logistic regression

Table 2 illustrates that univariate logistic analysis of the available data revealed statistically significant differences between the two groups in terms of occupation type, place of residence, operative mode, income, first-degree relatives, second-degree relatives, chronic disease status, periodontal disease status, history of tuberculosis, history of mumps, allergy history, dietary combination, breakfast habits, taste type, animal viscera, regular meals, dairy products, sweet, fish, drinking alcohol, damp dwelling, living above the second floor, dust exposure, and vaccine history, rs805297, rs9272219, HLA0405 ($P < 0.05$).

Table 3 illustrates that meaningful factors in the univariate logistic regression analysis were included as independent variables to fit the multivariate logistic regression model.

Occupation type (OR = 4.694, 95%CI = 2.375 ~ 9.278), place of residence (OR = 7.809, 95%CI = 2.916 ~ 20.911), history of mumps (OR = 6.809, 95%CI = 2.688 ~ 15.653), dietary combination (OR = 3.347, 95%CI = 1.327 ~ 8.446), damp dwelling (OR = 3.214, 95%CI = 1.166 ~ 8.860), history of vaccination (OR = 0.045, 95%CI = 0.017 ~ 0.123), rs805297 (OR = 0.363, 95%CI = 0.139 ~ 0.949), sweet (OR = 0.083, 95%CI = 0.015 ~ 0.462), and fish (OR = 6.206, 95%CI = 1.408 ~ 27.352) were significantly associated with RA.

Prediction model construction and assessment

On the basis of the results of the multivariate logistic regression analysis, a prediction model was built. The following formula was used:

Characteristics	The control group *n* (%)	The case group *n*(%)
Gender		
Female	194(50.0)	194(50.0)
Male	22(50.0)	22(50.0)
Ethnic group		
Han	210(50.5)	206(49.5)
Other	6(37.5)	10(62.5)
Educational level		
Primary school or less	56(55.4)	45(44.6)
Junior high	57(48.3)	61(51.7)
High school	22(38.6)	35(61.4)
College graduate	76(53.5)	66(46.5)
Postgraduate and above	5(35.7)	9(64.3)
Marital status		
Unmarried	34(63.0)	20(37.0)
Married or cohabiting	172(49.4)	176(50.6)
Divorce or separation	4(36.4)	7(63.6)
Bereaved	6(31.6)	13(68.4)
Place of residence		
Urban	188(56.6)	144(43.4)
Towns	13(37.1)	22(62.9)
Countryside	15(23.1)	50(76.9)
Occupation type		
Nonmanual labor	94(44.1)	119(55.9)
Manual labor	122(55.7)	97(44.3)
Operative mode		
On the job	165(63.0)	97(37.0)
Normal retirement	33(37.9)	54(62.1)
Retirement due to illness	2(6.2)	30(93.8)
Off the job	8(22.2)	28(77.8)
Current student	8(53.3)	7(46.7)
Medical payment		
Own expense	16(29.6)	38(70.4)
Employees' medical insurance	115(61.8)	71(38.2)
Resident medical insurance	84(44.4)	105(55.6)
Commercial insurance	1(33.3)	2(66.7)
Income		
≤ 1300	11(84.6)	2(15.4)
1301 ~ 2200	40(78.4)	11(21.6)
2201 ~ 3500	67(55.4)	54(44.6)
3501 ~ 5500	50(40.3)	74(59.7)
5501 ~ 6700	27(43.5)	35(56.5)
>6700	21(34.4)	40(65.6)

Table 1. Characteristics of RA patients and healthy individuals.

Y = logit*P*/(1-*P*)1.546 * nonmanual labor + 2.055 * countryside + 1.208 * partial meat - 2.493 * sweet3~4d - 1.472 * sweet1~2d - 0.832 * sweet < 1d + 1.825 * fish3~4d + 2.074 * fish5~7d + 1.167 * damp dwelling - 3.097 * history of vaccination + 1.870 * history of mumps -1.014* rs805297. According to the results of the multivariate logistic regression analysis, the nomogram function in the "rms" package in R was used to construct a nomogram (Fig. 1). Figure 2 showed that the AUC of this model was 0.912 (95% CI = 0.885–0.939). The Youden index was 0.699, the sensitivity was 0.847 and the specificity was 0.852. The Hosmer–Lemeshow test results ($\chi^2 = 8.441$, $P = 0.392$) indicated that the model had good diagnostic value.

Internal validation

Bootstrap validation was performed for internal validation of the model. Figure 3 showed that the internal validation AUC was 0.942 (95%CI = 0.922 ~ 0.962), which indicates that the model performs stably in the internal validation. The Hosmer–Lemeshow test results were ($\chi 2 = 8.147$, $P = 0.419$).

Variables	B	SE	Waldx2	*P*	OR	95%CI	
						Lower	Upper
Occupation type			44.721	<0.001			
Income			31.852	<0.001			
≤ 1300					1		
1301–2200	0.414	0.841	0.242	0.623	1.512	0.291	7.858
2201–3500	1.489	0.790	3.551	0.060	4.433	0.942	20.858
3501–5500	2.097	0.790	7.041	0.008	8.140	1.730	38.304
5501–6700	1.964	0.810	5.877	0.015	7.130	1.457	34.895
>6700	2.349	0.815	8.317	0.004	10.476	2.122	51.709
Operative mode			44.721	<0.001			
Current student					1		
On the job	1.024	0.255	16.075	<0.001	2.784	1.688	4.591
Normal retirement	3.239	0.741	19.088	<0.001	5.954	2.610	13.582
Retirement due to illness	1.784	0.421	17.973	<0.001	5.954	2.610	13.582
Off the job	0.398	0.533	0.556	0.456	1.488	0.524	4.232
Place of residence			24.660	<0.001			
Countryside					1		
Urban	−1.471	0.315	21.861	<0.001	0.230	0.124	0.426
Towns	−0.678	0.457	2.198	0.138	0.508	0.207	1.244
First-degree relative	0.815	0.263	9.567	0.002	2.259	1.348	3.785
Second-degree relative	0.731	0.334	4.799	0.028	2.078	1.080	3.998
Chronic disease status	0.764	0.278	7.559	0.006	2.147	1.245	3.700
Periodontal disease status	0.749	0.198	14.240	<0.001	2.114	1.433	3.119
History of tuberculosis	1.301	0.575	5.113	0.024	3.673	1.189	11.346
History of mumps	1.416	0.321	19.448	<0.001	4.122	2.197	7.737
Allergy history	−0.754	0.368	4.205	0.040	0.471	0.229	0.967
Dietary combination			18.262	<0.001			
Equilibrium					1		
Meat	1.011	0.304	11.063	<0.001	2.748	1.515	4.985
Vegetarian	0.828	0.255	10.513	0.001	2.289	1.387	3.775
Breakfast habits			8.668	0.013			
Everyday					1		
Sometimes	0.232	0.367	0.400	0.527	1.261	0.615	2.589
Always	0.710	0.242	8.629	0.003	2.035	1.267	3.269
Taste type			9.811	0.044			
Moderate					1		
Salted	−0.671	0.290	5.368	0.021	0.511	0.290	0.902
Light	−0.111	0.318	0.122	0.726	0.895	0.480	1.668
Sweet	−0.889	0.626	2.018	0.155	0.411	0.121	1.401
Fatty	−0.131	0.806	0.026	0.871	0.877	0.181	4.258
Animal viscera			11.881	0.036			
No					1		
>4/m	0.336	0.584	0.333	0.564	1.400	0.446	4.393
2–4/m	−1.101	0.344	10.266	0.001	0.332	0.170	0.652
1/m	−0.375	0.277	1.835	0.176	0.687	0.400	1.182
<1/m	−0.329	0.262	1.572	0.210	0.720	0.431	1.203
Regular meals	−1.144	0.319	12.864	<0.001	0.319	0.170	0.595
Dairy products			11.582	0.009			
<1 d					1		
5-7 d	−0.868	0.255	11.572	<0.001	0.420	0.255	0.692
3-4 d	−0.366	0.304	1.455	0.228	0.693	0.382	1.257
1-2 d	−0.348	0.255	1.857	0.173	0.706	0428	1.165
Sweet			12.904	0.005			
<1 d					1		
5-7 d	−1.761	0.569	9.587	0.002	0.172	0.056	0.524
3-4 d	−0.593	0.294	4.071	0.044	0.552	0.310	0.983
Continued							

Variables	B	SE	Waldx2	*P*	OR	95%CI	
						Lower	Upper
1-2 d	−0.338	0.225	2.266	0.132	0.713	0.459	1.108
Fish			23.970	<0.001			
<3 d					1		
3-4 d	1.780	0.556	10.232	0.001	5.930	1.992	17.650
5-7 d	2.380	0.548	18.857	<0.001	10.800	3.690	31.613
Drinking alcohol			8.912	0.012			
No					1		
Yes	−0.696	0.256	7.387	0.007	0.499	0.302	0.824
Quit	0.736	0.699	1.109	0.292	2.087	0.531	8.206
Damp dwelling	1.849	0.290	40.636	<0.001	6.355	3.599	11.220
Living above the second floor	−1.382	0.256	29.015	<0.001	0.251	0.152	0.415
Dust exposure	0.720	0.309	5.431	0.020	2.055	1.121	3.766
Vaccine history	−2.664	0.368	52.372	<0.001	0.070	0.034	0.143
rs805297	−0.674	0.319	4.471	0.034	0.510	0.273	0.952
rs9272219	−0.720	0.309	5.431	0.020	0.487	0.266	0.892
HLA0405	0.582	0.274	4.499	0.034	1.790	1.045	3.065

Table 2. Univariate logistic regression model for patients with RA.

Discussion

The analysis included 432 individuals, comprising 216 patients diagnosed with RA and an equal number of healthy patients. The mean age of patients in the case group was 48.27 ± 6.74 years, and that in the control group was 46.70 ± 6.33 years. In terms of sex, there were more women than men, which was basically consistent with the findings of the "China Rheumatoid Arthritis Development Report 2020"[41]. In this study, we developed a clinical prediction model for estimating RA incidence; thus, patients can be informed about the probability that they will develop RA. Multivariate logistic regression analysis revealed that occupational type, place of residence, mumps history, dietary combination, fish, sweet, damp dwelling, history of vaccination, and rs805297 were independent risk factors for RA.

Sociodemographic and environmental factors

This study showed that nonmanual labor is a risk factor for developing RA. This result was consistent with the results of a sampling survey on RA in Shanghai, China, which showed that the prevalence of RA was negatively correlated with labor intensity[42]. Although it has been suggested that nonmanual labor may be associated with an increased risk of developing RA, the results of related studies have been inconsistent[43,44]. Several studies have suggested that work stress may be associated with an increased risk of RA, and some people who are exposed to chronic stress may experience an imbalance in the regulation of the immune system, leading to an increased inflammatory response[45,46]. Inflammation is the underlying mechanism for RA and mood disorders[47]. The relationship between nonmanual labor and the immune system is complex, and current research has not reached consistent conclusions and warrants further study.

People who live in the countryside have a higher risk of RA. A study showed that the incidence of RA was greater in females living in rural areas (3.65/10,000 person-years) than in those living in urban areas (2.58/10,000 person-years)[49]. This may be because patients living in urban areas have easier access to better medical resources and a better understanding of disease prevention and treatment methods. On the other hand, the living environment in urban areas is more comfortable than that in rural areas. A total of 71.3% of the individuals in this study who lived in rural areas usually lived on the first floor, and the environment of the first floor of the rural areas was usually humid. Humidity, as a stimulating factor of the body's immune system, can induce or exacerbate the effects of several pathogenic factors and promote the occurrence of RA by causing a response from the immune system[49].

Genetic factors

Some studies suggest that the rs805297 allele is associated with RA risk, which is consistent with the results of this study. Among 520 RA patients, rs805297 was successful in 518 (99.6%) RA patients in a previous study[50]. Rs805297 is located on the promoter region of APOM (Apolipoprotein M) gene in the MHC III region on chromosome 6p21.3, and the MHC region is one of the loci closely related to RA. APOM is a high-density lipoprotein (HDL)-associated apoprotein, rs805297 polymorphism decreased its transcription activity slightly, which may affect RA susceptibility through altering the HDL level. Although APOM plays an important role in influencing immune function and thus can affect the risk of RA development, few studies have focused on the impact of APOM rs805297 on RA susceptibility[32,51]. So, more studies are needed to explore and analyze this in the future.

Variables	B	Waldχ^2	*P*	OR (95%CI)
Occupational type	1.546	19.785	<0.001	4.694 (2.375–9.278)
Place of residence		17.081	<0.001	
Urban	0.771	1.905	0.167	2.161 (0.724–6.456)
Towns	2.055	16.723	<0.001	7.809 (2.916–20.911)
History of mumps	1.870	17.304	<0.001	6.486 (2.688–15.653)
Dietary combination		7.325	0.026	
Partial meat	1.208	6.545	0.011	3.347 (1.327–8.446)
Partial vegetable	0.504	1.636	0.201	1.655 (0.765–3.581)
Sweet		15.526	0.001	
3-4 d	−2.493	8.070	0.005	0.083 (0.015–0.462)
1-2 d	−1.472	8.617	0.003	0.230 (0.086–0.613)
< 1 d	−0.832	5.701	0.017	0.435 (0.220–0.861)
Fish		8.123	0.017	
3-4 d	1.825	5.818	0.016	6.206 (1.408–27.352)
5-7 d	2.074	7.980	0.005	7.956 (1.887–33.545)
Damp dwelling	1.167	5.091	0.024	3.214 (1.166–8.860)
Vaccine history	−3.097	36.583	<0.001	0.045 (0.017–0.123)
rs805297	−1.014	4.273	0.039	0.363 (0.139–0.949)
rs9272219	−0.317	0.423	0.516	0.728 (0.280–1.894)
HLA0405	0.547	1.976	0.160	1.729 (0.806–3.708)
First-degree relative	0.671	2.994	0.084	1.956 (0.915–4.184)
Second-degree relative	0.813	0.482	0.092	2.254 (0.876–5.799)
Chronic disease status	−0.044	0.010	0.920	0.957 (0.405–2.259)
Periodontal disease status	0.425	1.980	0.159	1.529 (0.846–2.762)
History of tuberculosis	1.258	2.475	0.116	3.519 (0.734–16.873)
Allergy history	−0.853	2.151	0.143	0.426 (0.136–1.332)
Breakfast habits		2.582	0.275	
Meat	0.232	0.158	0.691	1.261 (0.403–3.947)
Vegetarian	0.622	2.581	0.108	1.863 (0.872–3.981)
Regular meals	−0.688	1.926	0.165	0.503 (0.190–1.328)
Dairy products		0.895	0.827	
5-7 d	−0.282	0.495	0.482	0.755 (0.344–1.653)
3-4 d	−0.060	0.015	0.903	0.941 (0.358–2.478)
1-2 d	−0.331	0.654	0.419	0.718 (0.322–1.603)
Taste type		2.109	0.716	
Salted	0.262	0.318	0.573	1.299 (0.523–3.229)
Light	0.633	1.630	0.202	1.884 (0.712–4.982)
Continued				

Variables	B	Waldχ^2	*P*	OR (95%CI)
Sweet	0.644	0.376	0.540	1.904 (0.243–14.944)
Fatty	0.006	0.000	0.996	1.006 (0.091–11.134)
Animal viscera		5.306	0.380	
>4/m	−1.455	2.202	0.138	0.233 (0.034–1.595)
2–4/m	−1.119	1.451	0.228	0.327 (0.053–2.017)
1/m	−1.038	1.284	0.257	0.354 (0.059–2.132)
<1/m	−21.524	0.000	0.999	0
Drinking alcohol		3.759	0.153	
Yes	0.564	0.319	0.572	1.758 (0.248–12.470)
Quit	−0.186	0.031	0.859	0.831 (0.107–6.446)
Living above the second floor	0.582	1.179	0.278	1.789 (0.626–5.113)
Dust exposure	−0.008	0.000	0.987	0.992 (0.392–2.512)

Table 3. Multivariate logistic regression model for patients with RA.

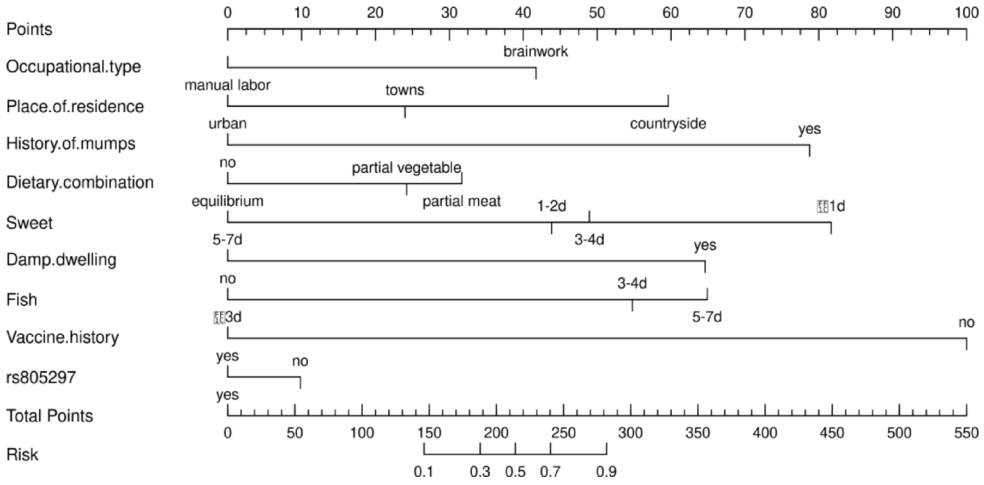

Fig. 1. RA prediction nomogram.

Disease history

The risk of RA in patients with a history of mumps was 2.426 times greater than that in patients without a history of mumps, which is similar to the results of many domestic and foreign studies. After an individual is infected with mumps, the mumps RNA virus stimulates Toll-like receptors in human plasmacytoid dendritic cells to produce interferon alpha, which further stimulates the body's immune system, resulting in autoimmunity[52]. For a long time, infection factors, including oral infection, lung infection, tuberculosis virus infection, EB virus infection, and tonsil infection, have been considered risk factors affecting the onset of RA[13]. The loss of self-tolerance can be caused by some mechanisms after the infection of pathogens, such as protein changes, molecular simulation, etc. At the same time, the RNA virus or DNA virus carried by the pathogen will stimulate the immune cells in the body to produce relevant inflammatory mediators, thereby inducing the occurrence of systemic autoimmunity[53]. Therefore, individuals at high risk of RA should pay attention to self-protection in daily life and try to avoid the occurrence of body infection caused by pathogens.

Dietary status

The risk of RA occurs in people who consumed a partial meat diet 1.608 times more than in those who ate a balanced diet. Taylor found that, compared with healthy people, RA patients have worse dietary intake, and the intake of high-fat food is a high risk factor for RA[54–57]. When the human body ingests meat, the body's digestive system will form small molecules that can be absorbed by the human body after digestion and decomposition of fats in the food, and some of these small molecules (such as α2-glycoprotein) may interact with the

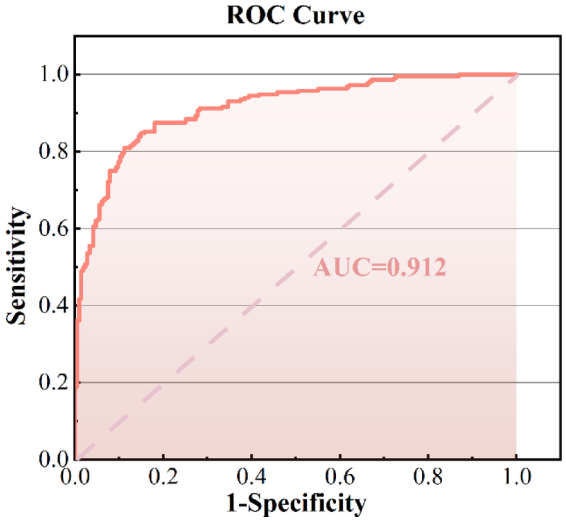

Fig. 2. AUC of the RA prediction model.

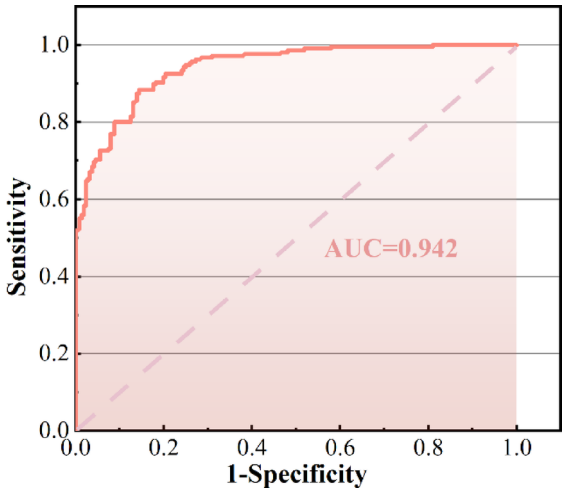

Fig. 3. AUC of the internal validation.

gastrointestinal epithelial barrier, mucosal immune system, and intestinal microflora, resulting in local and systemic inflammatory changes[55]. The intake of high-fat foods induces the production of an α2-glycoprotein, which stimulates lipolysis and fat loss, causing an increase in IL-17 in the immune system and leading to an inflammatory response in the body.

Sugars are most commonly reported to worsen RA. Studies have shown that excessive intake of dietary sugars can cause metabolic disorders and induce an increase in inflammatory mediators and certain proinflammatory cytokines in various tissues, leading to insulin resistance and low-grade chronic inflammation. Sugar is more likely to alter themicrobiome, thereby affecting downstream inflammatory pathways. Another way is that a high intake of sugars reduces Prevotella, which is an enteric bacteria associated with RA[58,59].

Fish was a risk factor for RA in this study, and Bottai showed that fish intake was positively correlated with RA risk[60]. However, a recent large prospective study of postmenopausal women did not find an association between the intake of omega-3 fatty acids from fish and RA. However, the relationship between fish intake and RA needs further investigation[61].

Lifestyle

Vaccination is a risk factor for RA. A comprehensive review revealed that rare autoimmune diseases, which include RA, may arise following vaccination[62]. However, the true incidence of these diseases after vaccination remains difficult to determine. The safety of vaccines has been proven, and these vaccines are an important means of preventing diseases.

Comparison with other models

The AUC of the RA risk prediction model constructed in this study was 0.912 (95% CI = 0.885–0.939), suggesting that the model was well differentiated and could distinguish the ability to develop RA. The performance of the prediction model in this study was better than that of the RA risk prediction model established by Chibni et al., with an AUC of 0.65[63]. This could be explained by the greater variety of predictors included in our study, which included both genetic risk factors and environmental factors. However, the model developed by Chibni and other scholars included only genetic loci as predictors. The performance of the prediction model in this study was better than that of the RA risk prediction model developed by Scott et al., who selected the British population as research objects in 2013. The predictive factors included genetic factors (HLA alleles and SNPs) and environmental factors (smoking), but the AUC) was 0.86 (95% CI, 0.86 ~ 0.91)[64]. This may be attributed to the fact that Scott and other researchers included only male smoking patients in the selection of research subjects, which introduced selection bias to some extent to affect the results.

Limitation

Although the model demonstrated favorable prediction performance and high prediction accuracy, several limitations should be acknowledged. Firstly, as a case-control study, this research may be subject to recall bias during data collection, and the absence of a prospective study design could impact the reliability of the findings. Secondly, due to constraints in time and funding, the study was limited to two hospitals and one community, resulting in the sample source being relatively lacking diversity and having a relatively small sample size. These limitations may restrict the generalizability of the model and increase the risk of overfitting. Therefore, future research should aim to optimize the model by integrating larger, more diverse, and representative datasets to further validate its robustness and reliability. Finally, this study was limited to internal validation and did not include external validation. Future research should conduct prospective external validation to enhance the model's applicability and accuracy across diverse populations and clinical settings.

Conclusion

In conclusion, the results of the RA incidence risk prediction model developed in this study showed that occupational type, place of residence, mumps history, dietary combination, fish, sweet, damp dwelling, history of vaccination, and rs805297 were independent risk factors for RA. Healthcare professionals should take preventive measures to reduce the risk of RA in this population. Individuals can reduce the risk of RA by reducing the intake of high-fat foods, avoiding prolonged exposure to humid environments, and preventing infections. The RA risk prediction model developed in this study has good performance, and the visualized graphs can be used by healthcare professionals as a more intuitive reference. This study provides strategies and support for screening and preventing RA in high-risk groups.

Data availability

The datasets are not publicly available due to them containing information that could compromise research participant privacy, the corresponding author may be contacted to obtain the data as reasonably necessary.

Received: 11 April 2024; Accepted: 30 April 2025
Published online: 13 May 2025

References

1. Smolen, J. S., Aletaha, D. & McInnes, I. B. Rheumatoid arthritis. *Lancet* **388**, 2023–2038 (2016).
2. An, J., Nyarko, E. & Hamad, M. A. Prevalence of comorbidities and their associations with health-related quality of life and healthcare expenditures in patients with rheumatoid arthritis. *Clin. Rheumatol.* **38**, 2717–2726 (2019).
3. Zeng, X. F., Zhu, S. L., Tan, A. C. & Xie, X. P. A systematic evaluation of rheumatoid arthritis disease burden and quality of life in China. A systematic evaluation of studies on disease burden and quality of survival of rheumatoid arthritis in China. *Chinese Journal of Evidence-Based Medicine* 13, 300–307 (2013). (2011).
4. Sokka, T., Abelson, B. & Pincus, T. Mortality in rheumatoid arthritis: 2008 update. *Clin. Exp. Rheumatol.* **26**, S35–61 (2008).
5. Hu, K. et al. Prevalence of rheumatoid arthritis among rural residents aged 20–79 years and risk factors in Qiannan, Guizhou. *China Public. Health*. **35**, 813–817 (2019).
6. Lo, J., Chan, L. & Flynn, S. A. Systematic review of the incidence, prevalence, costs, and activity and work limitations of amputation, osteoarthritis, rheumatoid arthritis, back pain, multiple sclerosis, spinal cord injury, stroke, and traumatic brain injury in the united States: A 2019 update. *Arch. Phys. Med. Rehabil.* **102**, 115–131 (2021).
7. Zhu, T. Y., Tam, L. S. & Li, E. K. Societal costs of rheumatoid arthritis in Hong Kong: a prevalence-based cost-of-illness study. *Rheumatol. (Oxford).* **50**, 1293–1301 (2011).
8. Deane, K. D. et al. Genetic and environmental risk factors for rheumatoid arthritis. *Best Pract. Res. Clin. Rheumatol.* **31**, 3–18 (2017).
9. Gerlag, D. M. et al. EULAR recommendations for terminology and research in individuals at risk of rheumatoid arthritis: report from the study group for risk factors for rheumatoid arthritis. *Ann. Rheum. Dis.* **71**, 638–641 (2012).
10. Kim, K., Bang, S. Y., Lee, H. S. & Bae, S. C. Update on the genetic architecture of rheumatoid arthritis. *Nat. Rev. Rheumatol.* **13**, 13–24 (2017).
11. Yamamoto, K., Okada, Y., Suzuki, A. & Kochi, Y. Genetics of rheumatoid arthritis in Asia–present and future. *Nat. Rev. Rheumatol.* **11**, 375–379 (2015).
12. Sun, J. Analysis of risk factors for rheumatoid arthritis and its prevalence in middle-aged and elderly population in Luohe area of Henan Province. *Chongqing Med.* **46**, 802–804 (2017).
13. Tobón, G. J., Youinou, P. & Saraux, A. The environment, geo-epidemiology, and autoimmune disease: rheumatoid arthritis. *Autoimmun. Rev.* **9**, A288–292 (2010).
14. Otsuki, T. et al. Immunological effects of silica and asbestos. *Cell. Mol. Immunol.* **4**, 261–268 (2007).
15. Alsaber, A. et al. Influence of ambient air pollution on rheumatoid arthritis disease activity score index. *Int. J. Environ. Res. Public. Health*. **17**, 416 (2020).

16. Jung, C. R., Hsieh, H. Y. & Hwang, B. F. Air pollution as a potential determinant of rheumatoid arthritis: A Population-based cohort study in Taiwan. *Epidemiology* **28** (Suppl 1), S54–S59 (2017).
17. van Beers-Tas, M. H., Turk, S. A. & van Schaardenburg, D. How does established rheumatoid arthritis develop, and are there possibilities for prevention? *Best Pract. Res. Clin. Rheumatol.* **29**, 527–542 (2015).
18. Kurreeman, F. et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am. J. Hum. Genet.* **88**, 57–69 (2011).
19. Sparks, J. A. & Costenbader, K. H. Genetics, environment, and gene-environment interactions in the development of systemic rheumatic diseases. *Rheum. Dis. Clin. North. Am.* **40**, 637–657 (2014).
20. Lahiri, M. et al. Using lifestyle factors to identify individuals at higher risk of inflammatory polyarthritis (results from the European prospective investigation of Cancer-Norfolk and the Norfolk arthritis Register–the EPIC-2-NOAR Study). *Ann. Rheum. Dis.* **73**, 219–226 (2014).
21. Hashimoto, M. et al. Periodontitis and Porphyromonas gingivalis in preclinical stage of arthritis patients. *PLoS One*. **10**, e0122121 (2015).
22. Lahiri, M., Morgan, C., Symmons, D. P. M. & Bruce, I. N. Modifiable risk factors for RA: prevention, better than cure? *Rheumatol. (Oxford)*. **51**, 499–512 (2012).
23. Alpízar-Rodríguez, D., Pluchino, N., Canny, G., Gabay, C. & Finckh, A. The role of female hormonal factors in the development of rheumatoid arthritis. *Rheumatol. (Oxford)*. **56**, 1254–1263 (2017).
24. Yarwood, A., Huizinga, T. W. J. & Worthington, J. The genetics of rheumatoid arthritis: risk and protection in different stages of the evolution of RA. *Rheumatol. (Oxford)*. **55**, 199–209 (2016).
25. van der Woude, D. Helm-van Mil, A. H. M. Update on the epidemiology, risk factors, and disease outcomes of rheumatoid arthritis. *Best Pract. Res. Clin. Rheumatol.* **32**, 174–187 (2018). van der.
26. de Hair, M. J. H. et al. Smoking and overweight determine the likelihood of developing rheumatoid arthritis. *Ann. Rheum. Dis.* **72**, 1654–1658 (2013).
27. Karlson, E. W. et al. Association of environmental and genetic factors and gene-environment interactions with risk of developing rheumatoid arthritis. *Arthritis Care Res. (Hoboken)*. **65**, 1147–1156 (2013).
28. Karlson, E. W. et al. Cumulative association of 22 genetic variants with seropositive rheumatoid arthritis risk. *Ann. Rheum. Dis.* **69**, 1077–1085 (2010).
29. Messemaker, T. C., Huizinga, T. W. & Kurreeman, F. Immunogenetics of rheumatoid arthritis: Understanding functional implications. *J. Autoimmun.* **64**, 74–81 (2015).
30. Rakieh, C. et al. Predicting the development of clinical arthritis in anti-CCP positive individuals with non-specific musculoskeletal symptoms: a prospective observational cohort study. *Ann. Rheum. Dis.* **74**, 1659–1666 (2015).
31. Scott, I. C. et al. Predicting the risk of rheumatoid arthritis and its age of onset through modelling genetic risk variants with smoking. *PLoS Genet.* **9**, e1003808 (2013).
32. Sparks, J. A. et al. Improved performance of epidemiologic and genetic risk models for rheumatoid arthritis serologic phenotypes using family history. *Ann. Rheum. Dis.* **74**, 1522–1529 (2015).
33. van de Stadt, L. A., Witte, B. I. & Bos, W. H. Schaardenburg, D. A prediction rule for the development of arthritis in seropositive arthralgia patients. *Ann. Rheum. Dis.* **72**, 1920–1926 (2013). van.
34. Bae, S. C. & Lee, Y. H. Alcohol intake and risk of rheumatoid arthritis: a Mendelian randomization study. *Z. Rheumatol.* **78**, 791–796 (2019).
35. Hozumi, H. et al. Acute exacerbation of rheumatoid arthritis-associated interstitial lung disease: mortality and its prediction model. *Respir Res.* **23**, 57 (2022).
36. van Boheemen, L. & van Schaardenburg, D. Predicting rheumatoid arthritis in At-risk individuals. *Clin. Ther.* **41**, 1286–1298 (2019).
37. Aletaha, D. et al. Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum* 62, 2569–2581 (2010). (2010).
38. Pattison, D. J. et al. Dietary risk factors for the development of inflammatory polyarthritis: evidence for a role of high level of red meat consumption. *Arthritis Rheum.* **50**, 3804–3812 (2004).
39. Dupont, W. D. Power calculations for matched case-control studies. *Biometrics* **44**, 1157–1168 (1988).
40. Internal validation of predictive models. Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).
41. Tian, X. P., Li, M. T. & Zeng, X. F. The challenges and opportunities for the management of rheumatoid arthritis in China: an annual report of 2019. *Zhonghua Nei Ke Za Zhi*. **60**, 593–598 (2021).
41. Shi, F. et al. Sample survey on the prevalence of arthritis in Shanghai and analysis of associated factors. *Chin. J. Epidemiol.* 72–76 (2003).
43. Syngle, D., Singh, A. & Verma, A. Impact of rheumatoid arthritis on work capacity impairment and its predictors. *Clin. Rheumatol.* **39**, 1101–1109 (2020).
44. Zhang, X. et al. The impact of rheumatoid arthritis on work capacity in Chinese patients: a cross-sectional study. *Rheumatol. (Oxford)*. **54**, 1478–1487 (2015).
45. Arleevskaya, M. et al. Interplay of environmental, individual and genetic factors in rheumatoid arthritis provocation. *Int. J. Mol. Sci.* **23**, 8140 (2022).
46. Malysheva, O., Pierer, M., Wagner, U. & Baerwald, C. G. O. [Stress and rheumatoid arthritis]. *Z. Rheumatol.* **69**, 539–543 (2010).
47. Pamukcu, M., İzci Duran, T., Ulusoy, H. & Altınbaş, K. Investigation of the correlation between mood disorder symptoms and disease activity and functional status in rheumatoid arthritis patients. *Turk. J. Med. Sci.* **51**, 3008–3016 (2021).
48. Yang, D. H., Huang, J. Y., Chiou, J. Y. & Wei, J. C.-C. Analysis of socioeconomic status in the patients with rheumatoid arthritis. *Int. J. Environ. Res. Public. Health*. **15**, 1194 (2018).
48. Ma, Y. et al. Correlation study between clinical symptoms and meteorological indicators in patients with rheumatoid arthritis with stable disease conditions. *Chin. J. Disease Control*. **19**, 904–908 (2015).
50. Huang, Y. et al. Apolipoprotein m (APOM) levels and APOM rs805297 G/T polymorphism are associated with increased risk of rheumatoid arthritis. *Joint Bone Spine*. **81**, 32–36 (2014).
51. Harrell, F. E. & Describing Resampling, Validating, and Simplifying the Model. in *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (ed. Harrell, Jr., Frank E.) 103–126Springer International Publishing, Cham, (2015). https://doi.org/10.1007/978-3-319-19425-7_5
52. Doria, A., Zampieri, S. & Sarzi-Puttini, P. Exploring the complex relationships between infections and autoimmunity. *Autoimmun. Rev.* **8**, 89–91 (2008).
53. Schmidt, R. E., Grimbacher, B. & Witte, T. Autoimmunity and primary immunodeficiency: two sides of the same coin? *Nat. Rev. Rheumatol.* **14**, 7–18 (2018).
54. Comee, L., Taylor, C. A., Nahikian-Nelms, M., Ganesan, L. P. & Krok-Schoen, J. L. Dietary patterns and nutrient intake of individuals with rheumatoid arthritis and osteoarthritis in the united States. *Nutrition* **67–68**, 110533 (2019).
55. Cutolo, M. & Nikiphorou, E. Don't neglect nutrition in rheumatoid arthritis! *RMD Open.* **4**, e000591 (2018).
56. Matsunaga, M., Lim, E., Davis, J. & Chen, J. J. Dietary quality associated with Self-Reported diabetes, osteoarthritis, and rheumatoid arthritis among younger and older US adults: A Cross-Sectional study using NHANES 2011–2016. *Nutrients* **13**, 545 (2021).

57. Na, H. S. et al. Th17 and IL-17 cause acceleration of inflammation and fat loss by inducing α2-Glycoprotein 1 (AZGP1) in rheumatoid arthritis with High-Fat diet. *Am. J. Pathol.* **187**, 1049–1058 (2017).
58. Ma, X. et al. Excessive intake of sugar: an accomplice of inflammation. *Front. Immunol.* **13**, 988481 (2022).
59. Tedeschi, S. K. et al. Diet and rheumatoid arthritis symptoms: survey results from a rheumatoid arthritis registry. *Arthritis Care Res. (Hoboken)*. **69**, 1920–1925 (2017).
60. Di Giuseppe, D., Wallin, A., Bottai, M., Askling, J. & Wolk, A. Long-term intake of dietary long-chain n-3 polyunsaturated fatty acids and risk of rheumatoid arthritis: a prospective cohort study of women. *Ann. Rheum. Dis.* **73**, 1949–1953 (2014).
61. Sparks, J. A. et al. Association of fish intake and smoking with risk of rheumatoid arthritis and age of onset: a prospective cohort study. *BMC Musculoskelet. Disord*. **20**, 2 (2019).
62. Guo, M., Liu, X., Chen, X. & Li, Q. Insights into new-onset autoimmune diseases after COVID-19 vaccination. *Autoimmun. Rev.* **22**, 103340 (2023).
63. Orellana, C. et al. Oral contraceptives, breastfeeding and the risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Ann. Rheum. Dis.* **76**, 1845–1852 (2017).
64. Seror, R. et al. Passive smoking in childhood increases the risk of developing rheumatoid arthritis. *Rheumatol. (Oxford)*. **58**, 1154–1162 (2019).

Acknowledgements
We gratefully acknowledge the assistance of the professionals from West China Hospital of Sichuan University, Sichuan Orthopedic Hospital, and Shi Yang Community Health Center.

Author contributions
L. T. and F. W. contributed to the study design. L. T., L. Q., X. L., and F. W.performed the data collection and curation. L. T. was involved in statistical analysis and wrote the manuscript. Y. S., H. H., Y. L., and H. C. reviewed and edited the manuscript.

Funding
The authors are grateful for the financial support from the West China Nursing Discipline Development Special Fund (HXHL20009).

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
This study was approved by the Ethics Review Committee and conducted in accordance with the Declaration of Helsinki. After review by the Ethics Committee on Biomedical Research, West China Hospital of Sichuan University, the study officially started with the ethical review number 2021 (175). The data used is considered anonymized. The study confirmed that informed consent was obtained from all subjects or their legal guardians.

Consent for publication
N/A (no identifiable information is published).

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00816-7.

Correspondence and requests for materials should be addressed to H.C.

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

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