A machine learning model for identifying systemic lupus erythematosus through laboratory information system and electronic medical record

J. Du[1], H. Huang[1], L. Pang[1], N. Duan[1], C. Huang[1], C. Liu[2], H. Li[1]

[1]Department of Clinical Laboratory, [2]Medical Records Statistics Office, Peking University First Hospital, Beijing, China.

Abstract

Objective

Systemic lupus erythematosus (SLE) is a heterogeneous autoimmune disease. Its diagnosis poses significant challenges especially at early stages and in atypical cases. The aim of this study was to develop a machine learning model based on common laboratory tests that can aid SLE diagnosis.

Methods

A standard protocol was developed to collect data of SLE and control immune diseases. A 10-fold cross-validation was performed in the modeling dataset (n=862), and an external dataset (n=198) was used for model validation. Machine learning algorithms were applied to construct a diagnostic model. Performance was evaluated based on area under the curve (AUC) values, F1-score, negative predictive value, positive predictive value, accuracy, sensitivity, and specificity.

Results

The optimal model was based on a random forest algorithm with 10 clinical features. Thrombin time, prothrombin activity, and uric acid contributed most to the diagnostic model. The SLE diagnostic model showed sufficient predictive accuracy, with AUC values of 0.8286 in the validation dataset.

Conclusion

Our diagnostic model based on 10 common laboratory tests identified the patients with SLE with high accuracy. An online version of the model can potentially be applied in clinical settings for the differential diagnosis of SLE.

Key words

systemic lupus erythematosus, diagnosis, laboratory information systems, machine learning, random forest

*Jialin Du, MD**
*Haiming Huang, MD**
Lu Pang, MD
Nan Duan, MD
Chenwei Huang, MD
Chenlong Liu, MD
Haixia Li, PhD

**These authors contributed equally and share first authorship.*

Please address correspondence to:
Haixia Li
Peking University First Hospital,
No. 8, Xishiku Street,
Xicheng District,
Beijing 100034, China
E-mail: bdyylhx@126.com

and to:
Chenlong Liu: chenlong.002@163.com

Received on March 28, 2023; accepted in revised form on September 25, 2023.

Funding: this study was supported by the National Natural Science Foundation of China (grant no. 82072369), the Scientific Research Seed Fund of Peking University First Hospital (grant no. 2021SF15), and the "Sailing Plan" of Medical Youth Science and Technology Innovation of Peking University (grant no. BMU2023YFJHPY003).

Competing interests: none declared.

Introduction

Systemic lupus erythematosus (SLE) is a complex autoimmune disease that can cause inflammation and injury in multiple organs, including skin, kidney, joints, nervous system, and blood elements (1). The active damage in the tissues and organs of patients with early SLE can be reversed by treatment, whereas chronic damage is often irreversible. Therefore, early diagnosis is an important factor that determines the prognosis of SLE (2). However, in clinical practice, SLE manifestations are extremely heterogeneous and multiple laboratory tests are needed for its diagnosis (3). Studies have found that patients with SLE often present with leukopenia, lymphopenia, and thrombocytopenia, with no features of musculoskeletal, skin, or other system involvement (4). Although anti-dsDNA and anti-Sm antibodies are specific markers of SLE, many patients with SLE lack these antibodies (5). To date, only a few biomarkers for SLE have been validated and used in clinical practice. The lack of pathognomonic features or tests poses a considerable challenge in SLE diagnosis (6). Moreover, professional equipment for detecting specific antibodies and complement is often not available in many medical institutions in China, including healthcare centers, community hospitals, and even some municipal hospitals. As a result, SLE diagnosis often relies on the acumen of physicians and requires a great deal of clinical experience when faced with complex clinical manifestations and limited laboratory results. In primary hospitals, the diagnosis of SLE may be delayed or initially missed if the index of suspicion is low. The 1997 American College of Rheumatology (ACR) criteria, the 2012 Systemic Lupus International Collaborating Clinics (SLICC) classification criteria, and the European League Against Rheumatism/American College of Rheumatology (EULAR/ACR) 2019 classification criteria are commonly used as diagnostic aids for SLE (7-9). Besides specific manifestations, the classification criteria include laboratory indexes that play critical roles in the diagnosis of SLE, including leukopenia, lymphopenia, thrombocytopenia, urine protein, pathological cast, and serum-specific antibodies. However, these criteria are not weighted for specificity, sensitivity, or disease severity, and therefore might exclude patients with early or limited SLE (10). A more efficient diagnostic tool is urgently required, particularly for differential diagnosis of suspected cases.

The use of big data in medicine has attracted growing and enthusiastic support in recent years (11). Machine learning (ML) has been widely applied in the medical field for disease diagnosis (12, 13), prediction (14, 15), and image recognition (16). These studies have shown that ML can assist clinicians in disease diagnosis by, for example, reducing the influence of subjective factors in the diagnosis process and improving the diagnostic efficiency by integrating clinical data. ML models have shown excellent pattern-recognising capability in the rheumatic immunology field, including SLE, and most of these models used complex clinical and laboratorial data as variables to diagnose SLE (17). Ma *et al*. (18) utilised the information From B cells and monocytes and established a ML model to distinguish SLE patients from healthy donors via not only scRNA-seq data but also bulk RNA-seq data. Cai *et al*. (19) employed deep learning to distinguish patients with SLE by skin imaging examination. Building robust ML models that avoid excessive complexity is still an important challenge. Although the increasing numbers of laboratory tests have played important roles in understanding SLE, there are still many laboratory tests that have not been adequately addressed. Lao *et al*. (20) showed that the neutrophil-to-lymphocyte ratio, red blood cell distribution width, and platelet-to-lymphocyte ratio were feature parameters that distinguished patients with SLE from healthy controls. Yang *et al*. (21) found that serum urea, creatinine, and uric acid were associated with skin rash, arthritis, erythrocytopenia, and thrombocytopenia in patients with SLE. However, the association between these clinically accessible markers and SLE remains unclear.

In this study, we developed an online diagnostic model based on ML methods using a new dataset in the Chinese pop-

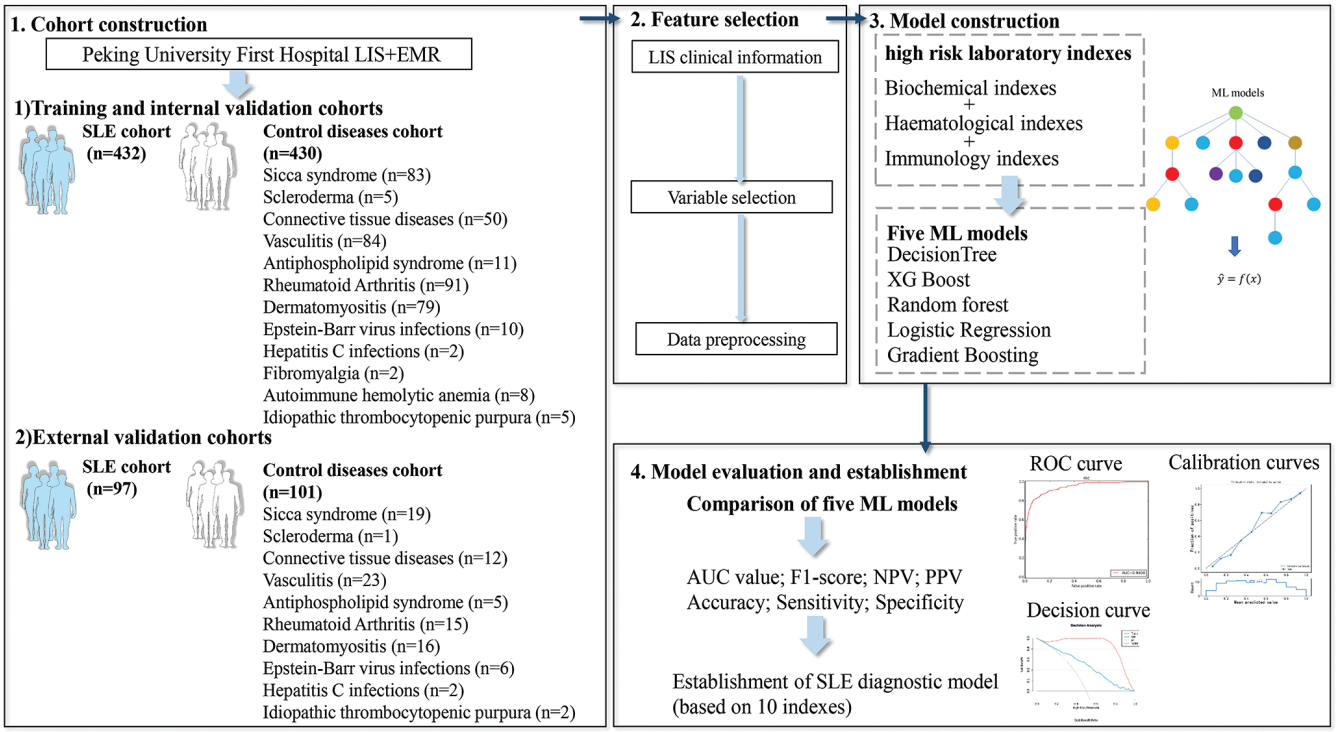

Fig. 1. A workflow to develop the SLE diagnostic model.
AUC: area under curve; NPV: negative predictive value; PPV: positive predictive value; ROC: receiver operating characteristic; SLE: systemic lupus erythematosus.

ulation to predict the patients at a high risk of SLE. The aim was to improve the diagnostic efficiency of SLE using objectively and accessible laboratory indexes, expand the capability of developing SLE diagnosis based on objective indicators, and eliminate the dependence on subjective clinical experience.

Materials and methods

Study population

We conducted a single-centre, retrospective study using the Laboratory Information System (LIS) database and Electronic Medical Records (EMR) database from Peking University First Hospital. We included patients diagnosed at Peking University First Hospital during 2008 and 2016 with SLE or miscellaneous control immune diseases that are relevant to the differential diagnosis of lupus. The disease control groups included patients with sicca syndrome, scleroderma, connective tissue diseases, vasculitis, antiphospholipid syndrome, antiphospholipid syndrome, dermatomyositis, Epstein-Barr virus infections, Hepatitis C infections, fibromyalgia, autoimmune haemolytic anaemia, and idiopathic thrombocytopenic purpura. Patients with SLE were identified according to the 1997 ACR criteria, 2012 SLICC classification criteria or 2019 EULAR/ACR classification criteria. The exclusion criteria were: 1) patients younger than 18 years old; 2) patients who were pregnant; 3) patients with two or more autoimmune diseases, such as patients with SLE and Sjögren's syndrome, scleroderma, antiphospholipid syndrome, rheumatoid arthritis, or connective tissue diseases, 4) patients with severe diseases including chronic cardiac insufficiency and liver diseases. According to the above inclusion and exclusion criteria, 1875 patients were selected from our hospital, among which 432 were patients with SLE and 1443 patients were patients with other immune diseases. Ultimately, a total of 432 SLE patients and 430 disease controls were included based on 1:1 propensity score matching (PSM) based on gender and age. An external test dataset was also collected of patients diagnosed at the Peking University First Hospital between 2017 and 2018 with SLE or control diseases to evaluate the performance of the ML model. This study was reviewed and approved by the Institutional Ethical Committee Board of Peking University First Hospital.

Data collection

The clinical parameters extracted from the LIS database and EMR database included demographic information, disease diagnoses, procedures (coded using ICD-10-CM) and laboratory tests. Baseline clinical and biochemical characteristics of patients were collected at their first visit. The extracted risk factors included: 1) immunology indexes, such as the immunoglobulins IgA, IgG and IgM; 2) haematologic indexes, such as white blood cell count, mean corpuscular haemoglobin concentration (MCHC), lymphocyte count, thrombin time (TT), prothrombin time (PT); and 3) biochemical indexes, such as 24-hour urine protein, uric acid (UA), urea, and lactate dehydrogenase (LDH).

Statistical analysis

A propensity score is a balancing score that can be used to account for the systematic differences between the exposure and control groups in an observational study. The PSM is applied so that the research subjects are comparable in clinical indicators for the purpose of balancing covariates and reducing bias. This method estimates the propensity score for each object with ranges of be-

Table I. Baseline clinical and biochemical characteristics of all patients.

Characteristic	SLE cohort (n=432)	Control diseases cohort (n=430)	*p* value
Age (years), median[IQR]	38 [29,50]	44 [34,49]	0.148
Gender			
Male	73 (16.9%)	73 (17.0%)	0.975
Female	359 (83.1%)	357 (83.0%)	
Laboratory test			
Biochemical indexes			
LDH (IU/L), median[IQR]	213.000 [166.000,279.427]	193.194 [155.000,279.000]	0.028*
TP (g/L), median[IQR]	67.300 [60.400,72.900]	68.700 [62.400,74.400]	0.005**
PA (mg/L), median[IQR]	225.000 [160.900,292.800]	216.600 [157.200,284.300]	0.155
Urea (mmol/L), median[IQR]	6.800 [4.840,11.170]	4.930 [3.790,6.800]	<0.001***
UA (μmol/L), median[IQR]	354.000 [272.000,456.000]	270.000 [211.000,338.000]	<0.001***
TCHO (mmol/L), median[IQR]	4.180 [3.510,5.230]	4.320 [3.770,5.160]	0.154
TBA (μmol/L), median[IQR]	4.100 [2.500,7.620]	4.400 [2.310,8.200]	0.425
TBIL (μmol/L), median[IQR]	9.500 [6.400,13.700]	9.600 [6.800,14.400]	0.407
DBIL (μmol/L), median[IQR]	1.360 [0.560,2.400]	1.207 [0.500,2.450]	0.736
ALP (IU/L), median[IQR]	70.000 [54.000,100.000]	73.000 [56.000,109.000]	0.185
CK (IU/L), median[IQR]	66.543 [41.000,94.000]	63.904 [43.772,103.105]	0.562
PCHE (IU/L), median[IQR]	7183.000 [5408.000,8816.000]	6831.000 [5576.000,8158.000]	0.093
ALT (IU/L), median[IQR]	16.000 [11.000,24.000]	15.000 [11.000,24.000]	0.338
GGT (IU/L), median[IQR]	20.000 [14.000,36.000]	20.000 [13.000,33.000]	0.714
AST (IU/L), median[IQR]	19.000 [15.000,25.185]	20.000 [16.000,25.000]	0.332
Cr (μmol/L), median[IQR]	78.912 [63.000,101.000]	80.000 [64.200,105.000]	0.596
HDL (mmol/L), median[IQR]	1.080 [0.890,1.360]	1.130 [0.900,1.380]	0.243
LDL (mmol/L), median[IQR]	2.560 [2.000,3.380]	2.470 [1.900,2.970]	0.003**
TG (mmol/L), median[IQR]	1.360 [0.950,2.110]	1.390 [0.920,1.990]	0.592
GLU (mmol/L), median[IQR]	5.140 [4.490,6.100]	5.330 [4.680,6.240]	0.007**
CL (mmol/L), median[IQR]	105.100 [102.300,107.600]	104.600 [102.400,107.000]	0.152
Ca (mmol/L), median[IQR]	2.250 [2.130,2.340]	2.240 [2.150,2.340]	0.739
P (mmol/L), median[IQR]	1.140 [1.010,1.330]	1.154 [1.020,1.370]	0.152
Na (mmol/L), median[IQR]	139.590 [137.700,141.460]	139.800 [137.300,141.600]	0.896
K (mmol/L), median[IQR]	3.840 [3.590,4.180]	3.870 [3.580,4.160]	0.908
Mg (mmol/L), median[IQR]	0.880 [0.820,0.950]	0.870 [0.807,0.950]	0.127
CRP (mg/L), median[IQR]	4.060 [1.990,9.664]	5.893 [2.333,16.300]	<0.001***
Haematologic indexes			
TT (s), median[IQR]	15.500 [14.700,16.819]	14.607 [13.965,15.533]	<0.001***
PT (s), median[IQR]	10.200 [9.600,10.900]	10.800 [10.200,11.800]	<0.001***
PTR, median[IQR]	0.990 [0.910,1.060]	1.050 [0.970,1.150]	<0.001***
PTA (%), median[IQR]	107.000 [95.000,120.000]	94.000 [84.000,107.000]	<0.001***
APTT (s), median[IQR]	31.994 [29.400,34.900]	31.400 [29.200,34.300]	0.378
APTTR, median[IQR]	1.080 [1.000,1.180]	1.070 [0.990,1.170]	0.897
INR, median[IQR]	0.990 [0.920,1.060]	1.050 [0.970,1.150]	<0.001***
D-D (mg/L), median[IQR]	0.400 [0.200,0.870]	0.270 [0.110,0.600]	<0.001***
WBC (10^9/L), median[IQR]	6.400 [5.000,8.800]	6.270 [5.000,8.092]	0.323
neutrophil count (10^9/L), median[IQR]	3.800 [2.440,5.930]	4.100 [2.490,6.270]	0.283
Monocyte count (10^9/L), median[IQR]	0.320 [0.200,0.500]	0.400 [0.280,0.520]	0.002**
lymphocyte count (10^9/L), median[IQR]	1.400 [0.900,1.960]	1.590 [1.060,2.120]	0.015*
RBC (10^9/L), median[IQR]	4.140 [3.520,4.600]	4.160 [3.560,4.590]	0.807
HGB (g/L), median[IQR]	125.485 [111.000,138.000]	123.000 [106.000,138.000]	0.192
HCT (%), median[IQR]	36.266 [32.600,40.500]	35.845 [31.400,39.300]	0.104
MCH (pg), median[IQR]	30.200 [28.800,31.500]	30.300 [28.500,31.500]	0.607
MCHC (g/L), median[IQR]	346.000 [336.000,354.000]	343.000 [333.000,352.000]	0.009**
MCV (fl), median[IQR]	87.100 [83.200,91.500]	87.600 [83.500,91.700]	0.631
PLT (10^9/L), median[IQR]	216.000 [160.000,280.000]	209.000 [155.000,263.000]	0.231
PCT (%), median[IQR]	0.249 [0.190,0.295]	0.245 [0.202,0.295]	0.515
RDW (%), median[IQR]	14.000 [13.100,15.300]	13.700 [12.900,15.300]	0.063
ESR (mm/1h), median[IQR]	22.000 [10.000,49.000]	19.000 [10.000,40.000]	0.041*
Immunology indexes			
lgG (g/L), median[IQR]	12.800 [9.050,17.800]	13.900 [10.465,18.000]	0.012*
lgM (g/L) ,median[IQR]	0.982 [0.590,1.460]	1.340 [0.900,1.831]	<0.001***
lgA (g/L), median[IQR]	2.624 [1.820,3.520]	2.421 [1.700,3.360]	0.068

Characteristic	SLE cohort (n=432)	Control diseases cohort (n=430)	*p* value
Urine indexes			
24hrUpr (g/24h), median[IQR]	1.544 [0.338,4.120]	1.340 [0.378,3.236]	0.075
Pathological cast (/LP), median[IQR]	0.149 [0.000,0.800]	0.100 [0.000,0.500]	0.017
U-WBC (/μl), median[IQR]	10.000 [3.600,31.900]	7.300 [2.900,21.700]	0.02**
U-RBC (/μl), median[IQR]	11.200 [4.700,38.900]	8.914 [3.500,31.700]	0.023**
U-GLU			
-	398 (92.13%)	390 (90.698%)	0.863
1/2+	11 (2.546%)	13 (3.023%)	
+	5 (1.157%)	9 (2.093%)	
++	7 (1.62%)	5 (1.163%)	
+++	4 (0.926%)	5 (1.163%)	
++++	7 (1.62%)	8 (1.86%)	
U-URO			
-	417 (96.528%)	397 (92.326%)	0.052
+	12 (2.778%)	23 (5.349%)	
++	1 (0.231%)	8 (1.86%)	
+++	1 (0.231%)	1 (0.233%)	
++++	1 (0.231%)	1 (0.233%)	

Values are presented as median (IQR) for continuous variables or n (%) for binary variables, * $p<0.05$; ** $p<0.01$; *** $p<0.001$.
24hrUpr: 24-hour urine protein; ALP: alkaline phosphatase; ALT: alanine aminotransferase; APTT: activated partial thromboplastin time; APTTR: activated partial thromboplastin time ratio; AST: aspartate amino transferase; Ca: calcium; CK: creatine kinase; CL: chlorine; Cr: creatinine; CRP: C-reactive protein; DBIL: direct bilirubin; D-D: d dimer; ESR: erythrocyte sedimentation rate; GLU: glucose; GGT: gamma-glutamyl transpeptidase; HCT: haematocrit; HDL: high-density lipoprotein; HGB: haemoglobin concentration; INR: international normalised ratio; K: kalium; LDH: lactic dehydrogenase; LDL: low density lipoprotein; lgA: immunoglobulin A; lgG: immunoglobulin G; lgM: immunoglobulin M; MCH: mean corpuscular haemoglobin content; MCHC: mean corpuscular haemoglobin concentration; MCV: mean corpuscular volume; Mg: magnesium; Na: natrium; P: phosphorus; PA: prealbumin; PCHE: cholinesterase; PCT: platelet haematocrit; PLT: blood platelet count; PT: prothrombin time; PTA: prothrombin activity; PTR: prothrombin time ratio; RBC: red blood cell; RDW: red blood cell distribution width; SD: standard deviation; TBA: total bile acid; TBIL: total bilirubin; TCHO: total cholesterol; TG: triglyceride; TP: total protein; TT: thrombin time; UA: uric acid; U-GLU: urine glucose; U-RBC: urinary red cell count; U-URO: urine urobilinogen; U-WBC: urinary white blood cell count; WBC: white blood cell.

tween 0 and 1, representing the probability of the subject being classified to the treatment group. It discards the subjects who are not matched thus resulted in smaller sample size. Therefore, PSM was applied to balance the distribution of covariates (age and gender) between the SLE and control group.

The raw dataset had some noises (contained errors, outlier values), missing and inconsistency values that could reduce the quality of our dataset and affected the model's performance. Therefore, a feature selection process was adopted before model construction. The data preprocessing procedure were carried out as follows, 1) Removed variables with >30% missing value; 2) Removed outlier values; 3) K-nearest neighbors method was used for missing data imputation. Finally, we kept 60 potential features after consulting extensive literature search and discussing with expert in this area. Continuous variables were presented as mean ± standard deviation (SD) for normally distributed variables or median with interquartile range (IQR) for non-normally distributed variables, and categorical variables were presented as percentage frequencies. Demographic and laboratory tests were compared using a t-test or Mann-Whitney U-test for continuous and chi-square test for categorical variables. We applied Kolmogorov-Smirnov Normality test to test for normality, and Levene's-test to test homogeneity of variances. Then, normally distributed variables were compared by the Student's t-test, and non-normally distributed variables were compared by Mann-Whitney U-test. All statistical tests were two-tailed and $p<0.05$ was considered significant. SPSS (v. 25.0), R (v. 3.6.1), and Python (v. 3.4.3) were were systematically used for statistical analysis. Model construction and visualisation were carried out using Deepwise and Beckman Coulter DxAI platform.

Model construction and evaluation

In the training cohort, least absolute shrinkage and selection operator (LASSO) logistic regression analysis was utilised to rank the importance of risk factors. In LASSO regression, the beta coefficients of variables that are not strongly associated with the outcome are decreased to zero, which removed these variables from the model. Ten Features were confirmed by the LASSO regression and were further selected into the ML model construction. 10-fold cross-validation was applied to the modeling dataset, using 9 of the folds as the training set to train the model, and the remaining 1-fold as the internal validation to score the model. Five ML models were constructed to predict the occurrence of SLE. The five models are Decision tree, XGBoost, Random forest, Logistic regression, gradient boosting. The detailed information of the five ML models is as follows:1) Decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences; 2) XGBoost is an implementation of gradient boosted decision trees designed for better speed and performance; 3) Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time; 4) Logistic Regression applies the logistic function to predict the probability of the class in

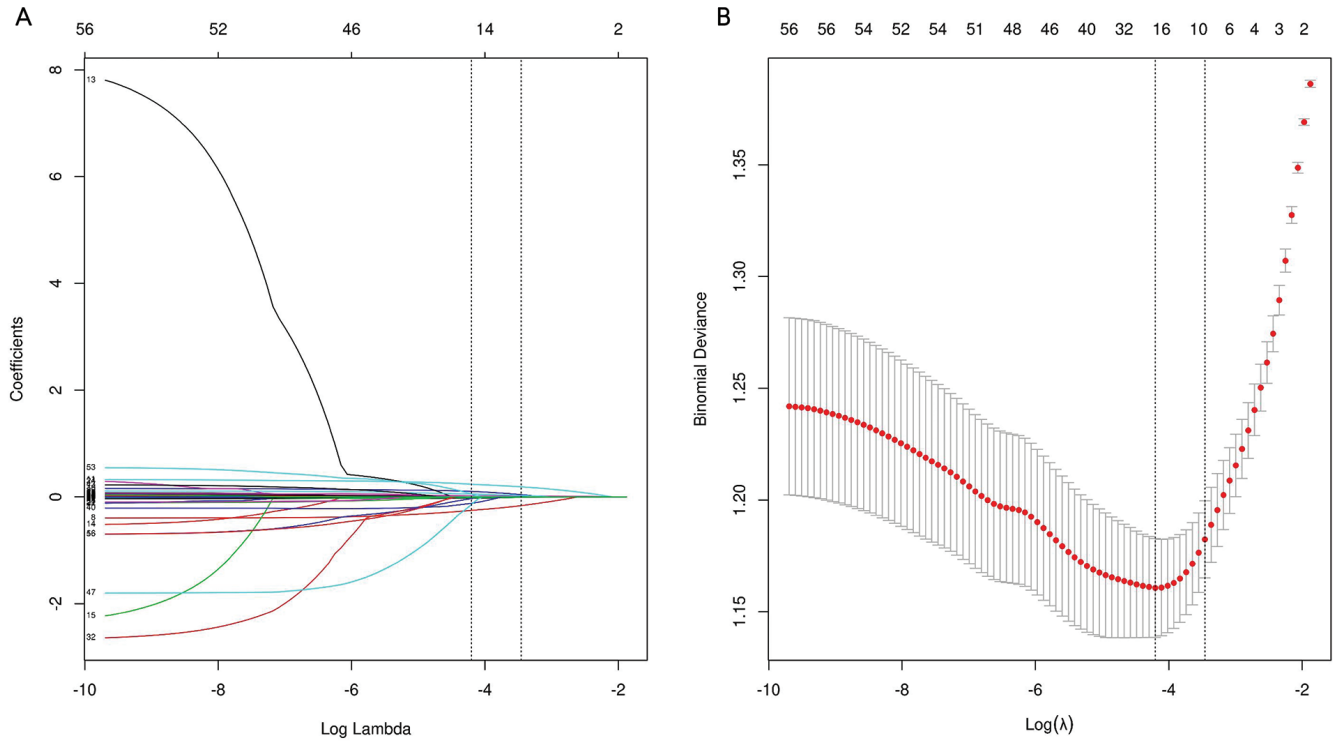

Fig. 2. LASSO regression analysis.
A: 10-fold cross-validation was used to draw vertical lines at selected values, where the optimal lambda produces ten nonzero coefficients.
B: Tuning parameter (lambda) selection cross-validation error curve.

Table II. Comparative analysis of the model performance.

	ML model	AUC	F1-score	NPV	PPV	Accuracy	Sensitivity	Specificity
Training cohort	DecisionTree	0.8299	0.716	0.7002	0.8338	0.7506	0.6273	0.8744
	XGBoost	0.9609	0.8978	0.8871	0.9117	0.8991	0.8843	0.914
	RandomForest	1	1	1	1	1	1	1
	LogisticRegression	0.788	0.7053	0.7037	0.707	0.7053	0.7037	0.707
	GradientBoosting	0.9186	0.8228	0.8102	0.846	0.8271	0.8009	0.8535
Internal validation cohort	DecisionTree	0.7258	0.6273	0.6372	0.7242	0.6705	0.5532	0.7884
	XGBoost	0.7999	0.7124	0.7082	0.7288	0.7181	0.6968	0.7395
	RandomForest	0.8286	0.7568	0.7517	0.7685	0.7599	0.7454	0.7744
	LogisticRegression	0.775	0.6952	0.6939	0.6935	0.6937	0.6968	0.6907
	GradientBoosting	0.8066	0.7292	0.7235	0.7488	0.7355	0.7106	0.7605

AUC: area under curve; ML: machine learning; NPV: negative predictive value; PPV: positive predictive value

a two-class problem. It is often used to predict the risk of developing a given disease; 5) Gradient Boosting is an ensemble of weak prediction models and minimises the loss function by adding weak learners using gradient descent. To evaluate and compare the performances of the five ML models, a receiver operating characteristic (ROC) curve was constructed and areas under the ROC curve (AUCs) with 95% confidence intervals were calculated. Five measurement criteria (F1-score, sensitivity, specificity, positive prediction value (PPV), negative prediction value (NPV)) were calculated and compared to select the best ML model. Furthermore, the calibration curve was used to assess the agreement between the prediction probabilities and the sample probabilities; and the decision curve analysis (DCA) was used to assess the clinical benefit of the model. The interpretation of the model is performed by SHAP, which calculated the contribution and influence of each feature toward the final prediction precisely. The SHAP values can show how much each predictor contributes, either positively or negatively to the outcome variable. The workflow used to develop the ML model for SLE is shown in Figure 1

Results

Baseline characteristics

After PSM, 432 patients with clinically diagnosed with SLE and 430 control patients with other immune diseases groups were selected remained after the first step of the feature selection process. Before PSM, a significance difference between age and gender were observed between SLE and disease control groups ($p<0.001$, $p<0.001$, Supplementary Table S1). Baseline de-

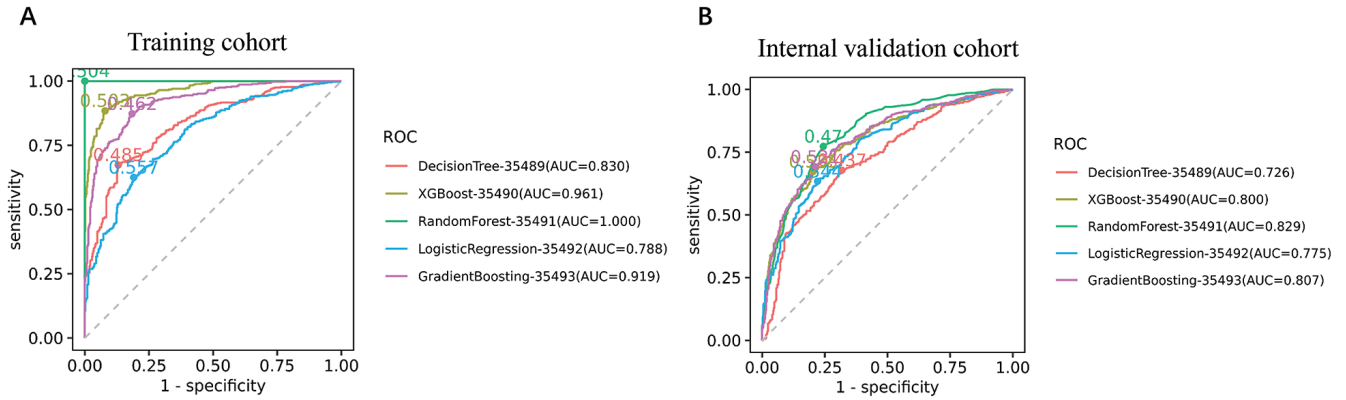

Fig. 3. The ROC curves show the discriminative ability of the five ML models.
A: The AUC in the training cohort; **B**: The AUC in the internal validation cohort. AUC: area under curve; ROC: receiver operating characteristic.

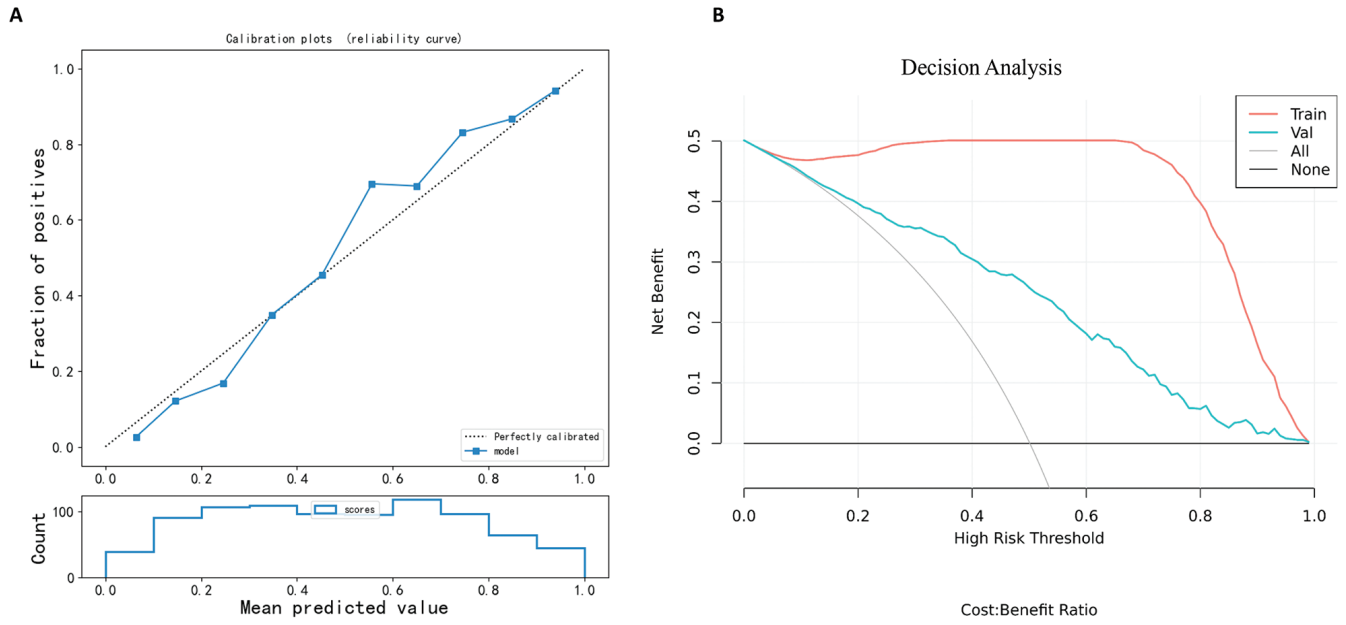

Fig. 4. Evaluation of validity and reliability of the random forest model.
A: Calibration curve analysis of the internal validation set. **B**: Decision curve analysis of the training set and the internal validation set.

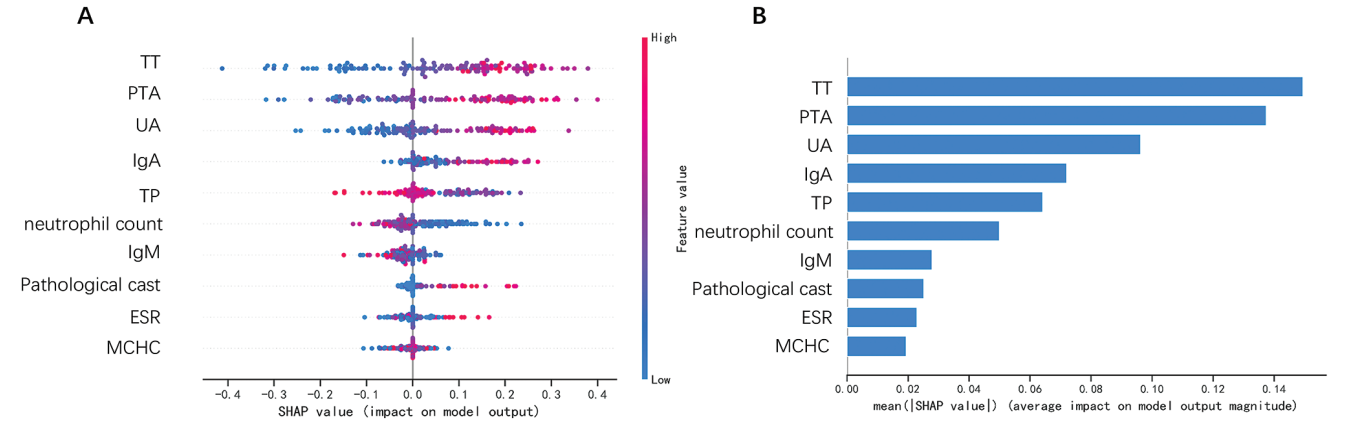

Fig. 5. The SHAP to Model Interpretation (**A**) The SHapley Additive exPlanation (SHAP) values. Redder sample points indicate the value of the feature is larger, and bluer sample points indicate the value of the feature is smaller.
B: The weight of variable importance as indicated by SHAP. The matrixdiagram describes the importance of each covariate in the development of the final diagnostic model.
ESR: erythrocyte sedimentation rate; IgA: immunoglobulin A; IgM: immunoglobulin M; MCHC: mean corpuscular haemoglobin concentration; PTA: prothrombin activity; TP: total protein; TT: thrombin time; UA: uric acid.

Table III. The performance of the model in the external validation cohort.

RF model	AUC	Sensitivity	Specificity
External validation cohort	0.706	0.607	0.708

AUC: area under curve; RF: random forest.

mographic and laboratory test features of the patients in the SLE and control cohorts after PSM are summarised in Table I, where age and gender were well balanced between the two groups. The median ages of the patients in the SLE and control cohorts were 38 (29, 50) and 44 (34, 49) years old, respectively. The percentage of females in the two cohorts was almost five times that of males. Twenty two of the laboratory tests showed significant differences between the two cohorts, namely the biochemical indexes LDH, total protein (TP), low density lipoprotein, C reactive protein, UA, urea and glucose; the haematology indexes TT, PT, PT ratio, PT activity (PTA), international normalised ratio, D dimer, monocyte count, lymphocyte count, MCHC and erythrocyte sedimentation rate (ESR); the immunology indexes IgG and IgM; the urine index pathological cast, urinary white blood cell count and urinary red blood cell count .

ML model establishment and evaluation

After LASSO regularisation (lambda with minimum mean square error de = 0.031), 10 clinical features, namely biochemical indexes (UA, TP), immunology indexes (IgA, IgM), haematologic indexes (TT, PTA, neutrophil count, ESR, MCHC), and urine index (Pathological cast), were included in the algorithm. The coefficients are shown in Supplementary Table S2, and a coefficient profile is plotted in Figure 2A. A cross-validated error plot of the LASSO regression model is shown in Figure 2B.

To explore the optimal diagnostic model, we compared five commonly used ML algorithms, Decision Tree, XGBoost, Random forest, Logistic Regression and Gradient Boosting. Comparatively, RF algorithm had the highest predictive performance among the five models (Table II) in both training cohort and internal validation cohort. The AUC value, F1-score, NPV, PPV, accuracy, sensitivity and specificity of RF model was 1, 1, 1, 1, 1, 1 and 1 respectively in training cohort. The performance of this model was slightly decreased in internal validation cohort, which had an AUC of 0.8286, F1-score of 0.7568, NPV of 0.7517, PPV of 0.7685, accuracy of 0.7599, sensitivity of 0.7454 and specificity of 0.7744. The ROC curves of each ML model were shown in Figure 3.

Explanation of Random Forest Model

The ROC curve shows that the random forest model had good classification ability in predicting the risk of SLE (AUC = 0.8286, sensitivity = 0.7454, specificity = 0.7744). The calibration cure showed that the model's predicted probabilities were in good agreement with the actual probabilities (Fig. 4A); and decision curve analysis (DCA) indicated that the model had high clinical benefits (Fig. 4B). These aforementioned results indicated that the RF model was well-fitted and accurately diagnose SLE risks. The visualisation of the diagnostic model was displayed online through Deepwise and Beckman Coulter DxAI platform.

To detect the positive and negative relationships of the predictors with SLE, SHAP values were applied to uncover the impact of the risk factors. The 10 most important features selected by random forest are shown in Figure 5A. In each feature important line, the attributions of all patients to the results are plotted with different coloured dots, where blue dots represent low risk values and red dots represent high risk values. Compared with the control immune diseases, increased TT, PTA, UA, IgA, pathological cast, and ESR, and decreased TP, IgM, neutrophil count, and MCHC contributed to the diagnosis of SLE. The ranking of the 10 features evaluated by the average absolute SHAP value is shown in Figure 5B. The SHAP value on the x-axis indicates the importance of the diagnosis model.

External validation study

We collected another 198 cases as an external test dataset to further evaluate the performance of the SLE diagnostic model; 97 were patients with SLE and 101 were patients with the control diseases. Baseline characteristics of all patients are summarised in Supplementary Table S1. The diagnostic performance of our model is shown in Table III. The AUC value, sensitivity, and specificity were 0.706, 0.607, and 0.708 respectively. These results indicate that the constructed ML diagnostic model based on laboratory test results had comparable diagnostic ability.

Discussion

The diagnosis of SLE in clinical practice is challenging and depends on the clinical experience and expertise of rheumatologists. The clinical manifestations of SLE are atypical and insidious, and share symptoms in common with other diseases. Accurate diagnosis of SLE is challenging and often leads to delayed diagnosis (2, 22). Furthermore, multiple laboratory tests are performed in SLE diagnosis, and instruments that are available in primary hospitals are often limited, leading to misdiagnosis and missed diagnosis. In this study, we used an RF algorithm to develop a diagnostic model that could distinguish patients with SLE from patients who did not have SLE based on 10 common laboratory indicators that cover most patients in areas where there are only limited healthcare resources.

ML and data-driven approaches are becoming very important, especially in the medical field. These approaches address traditional limitations by using underlying connections that cannot be discovered with other statistical techniques to make accurate decisions ML analysis is particularly useful for research in complex chronic diseases, such as rheumatic autoimmune inflammatory diseases, in which the disease conditions are extremely heterogeneous and multiple factors contribute to disease diagnosis and progression.

Other studies have established ML algorithms to classify patients with SLE using combinations of multiple indicators. In a Swedish study, an RF classifier with AUC of 0.78 was built to classify patients with SLE using genotype data (23). Ceccarelli *et al*. (24) incorporated demographic data and laboratory and clinical parameters, and used an artificial neural network model to identify risk of chronic organ damage in SLE. Maffi *et al*. (25) established ML techniques that correctly predicted difficult-to-treat flares based on baseline clinical variables. Such approaches may help to support clinicians in their treatment decisions. However, less research has been directed towards the diagnosis of SLE, which needs to be differentiated from rheumatoid arthritis, myositis, sicca syndrome, connective tissue diseases, and other immune diseases (26, 27). No single biomarker can be sensitive and specific enough for SLE, and therefore combinations of multiple biomarkers are needed to help clinicians make comprehensive judgements. ML potentially has the utility and power in this context. Several studies have analysed blood polypeptides and lipids and distinguished patients with SLE from control groups using ML approaches (28-30). Adamichou *et al*. (31) developed an accurate algorithm based on classical disease features that can aid SLE diagnosis and assess severe forms. Their model included clinical features that require subjective judgement and specific antibodies and complement that were not widely available in primary hospitals. Although all of these studies yielded useful results, detecting some of the indicators included in the model is complicated and the clinical application is limited. The diagnostic ML model that we built showed high predictive ability for SLE, and had good discriminative ability in predicting patients with SLE in both the internal validation and external validation cohorts. The model based on 10 factors performed well, with AUC values of 1, 0.8286, and 0.706 in the training, internal, and external validation sets, respectively.

The online diagnostic model built in this study will enable clinicians to identify patients with SLE based on objective laboratory indicator values using portable laptops or mobile devices. The model includes only 10 common laboratory indexes that are more clinical accessibility and less costly than other SLE biomarkers that have been used, such as autoantibodies and inflammation factors. Our model also performed better than previous models, especially in identifying patients with SLE who did not have typical clinical symptoms or lacked specific serological features. For healthcare centres, community hospitals, and even some municipal hospitals in China, the available of clinical laboratory tests are inadequate, and therefore our model can be used to help physicians identify patients with SLE and distinguish them from patients with other immune diseases based on common laboratory indicators.

The LASSO screened and ML model constructed in this study identified 10 risk factors with the highest explanatory power for SLE diagnosis, namely TT, PTA, UA, IgA, TP, IgM, neutrophil count, pathological cast, ESR, and MCHC. Five of these features are included in the SLICC criteria (7), namely IgA, IgM, neutrophil count, pathological cast, and MCHC. Additionally, some new laboratory tests from the SLICC criteria were identified as predictors, such as TT, PTA, UA, TP, and ESR, indicating their contribution to the disease network may provide clues for a deeper understanding of the pathogenesis of SLE. Several studies have reported laboratory indexes as the clinical presentation for patients with SLE.

In our study, the SHAP plot shows TT as the largest contributor to the model prediction, indicating its important role in the diagnosis of SLE. In practice, thrombotic complications and coagulation disorders contribute significantly to morbidity and mortality rates in patients with SLE (32). Compared with the control diseases group, the patients with SLE have abundant immune complexes and activated complement system in the blood, which likely leads to platelet activation (33). Phosphorylated fibrinogen is generated by activated platelets, leading to an increase of its coagulability and promote thrombotic complications in SLE patients (34). A study reported thrombocytopenia has a high prevalence in SLE patients and is related to increased TT (35). PTA was also accounted for a high weight in the ML model, which was calculated from PT and is a significant coagulation biomarker that reflect the clotting ability of blood. Previous studies have shown that anti prothrombin antibodies could contribute to thromboembolic risk assessment and stratification of patients with SLE, which may affect the test results of PT and PTA (36). Fujiwara *et al*. emphasized that measuring the PT might be required in patients with Lupus anticoagulant-hypoprothrombinaemia syndrome when they do not have a typical clinical course or distinctive symptoms (37). These coagulation indicators are not mentioned specifically in the classification criteria of SLE. Our results highlight the important relationship between SLE and coagulation and provide some ideas for the diagnosis and treatment of SLE in clinical practice.

Lupus nephritis is a common manifestation of SLE that arises as a result of antibody–antigen complexes that deposit in the glomerulus and cause a thickening of the basement membrane (32, 33, 38). In this study, UA and Pathological cast, the renal dysfunction marker, were selected into the diagnostic model. A study by Yang *et al*. (21) found that increased amounts of UA accompanied by erythrocytopenia had an independent positive association with thrombocytopenia and negative association with skin rash and arthritis in patients with SLE. The antibody binding to multiple intrarenal autoantigens induced more obvious tubular injuries and results in pathological cast formation and tubular dilatation (39). Because of renal damage and the production of proteinuria, TP was decreased in patients with SLE (40).

Haematological abnormalities are very frequently associated with SLE (4). For haematologic indexes, MCHC, Neutrophil count and ESR were included in ML model. Anaemia is particularly common in patients with SLE (41), and MCHC was included in the ML model as an indicator for the diagnosis of anaemia (42). Several studies have shown that changes in iron homeosta-

sis, impaired erythropoiesis and decreased EPO secretion were associated with decreased MCHC level in patients with SLE (43). In patients with SLE, thrombocytopenia and autoimmune aetiology contributes directly to neutropenia (44). ESR have been reported as useful reliable markers for assessing immune, inflammatory response, and disease activity in patients with SLE (45). Other indicators associated with immune dysfunction, such as IgA and IgM, have been shown to be associated with the pathogenesis of SLE that characterised by the production of autoantibodies to a broad range of self-antigens (46). We applied ML on panels of these laboratory indexes to construct a model that can distinguish SLE from competing rheumatologic conditions.

Our study has several limitations. First, clinical symptoms and social determinations could potentially be useful for the development of an ML model for SLE, but these data were not available in the LIS. Second, the prediction model was based on clinical information on patients' first visit, but might be affected by different courses and treatment of the disease. Therefore, a prospective clinical study and a subgroup analysis of patients with the same courses could be carried out in the future to minimise bias. Third, the patients with SLE were from a single centre research, and therefore, the number of patients was limited. Other centres, other populations, and more clinical features should be included in future studies to improve and verify our prediction model. Fourth, we did not implement our model in real clinical practice, and therefore the clinical value of our model remains unknown.

Overall, this study showed the potential of diagnosis of atypical SLE based on common laboratory indexes, which could help physicians fast screen patients for SLE even with limited resources or experience.

Conclusions

We created an online portable model for predicting SLE based on LIS and EMR information that does not rely on the clinical experience of rheumatologists and can accurately diagnosis SLE with objective and easily accessible laboratory tests. Such a model will be valuable for improving the efficiency of screening patients with suspected SLE and will provide an accessible tool for primary care clinicians with limited healthcare resources. Moreover, the 10 laboratory indexes screened by the ML model provide a new idea and reference for SLE pathogenesis research.

Acknowledgments

The authors thank Hongyun Yang for her assistance with data sorting.

References

1. CONNELLY K, MORAND EF: Systemic lupus erythematosus: a clinical update. *Intern Med J* 2021; 51(8): 1219-28. https://doi.org/10.1111/imj.15448
2. KIRIAKIDOU M, CHING CL: Systemic lupus erythematosus. *Ann Intern Med* 2020; 172(11): Itc81-itc96 https://doi.org/10.7326/aitc202006020
3. TSOKOS GC: Systemic lupus erythematosus. *N Engl J Med* 2011; 365(22): 2110-21. https://doi.org/10.1056/nejmra1100359
4. SASIDHARAN PK, BINDYA M, SAJEETH KUMAR KG: Hematological manifestations of SLE at initial presentation: is it underestimated? *ISRN Hematology* 2012; 2012: 961872. https://doi.org/10.5402/2012/961872
5. PISETSKY DS: Anti-DNA antibodies - quintessential biomarkers of SLE. *Nat Rev Rheumatol* 2016; 12(2): 102-10. https://doi.org/10.1038/nrrheum.2015.151
6. ZUCCHI D, ELEFANTE E, SCHILIRÒ D *et al*.: One year in review 2022: systemic lupus erythematosus. *Clin Exp Rheumatol* 2022; 40(1): 4-14. https://doi.org/10.55563/clinexprheumatol/nolysy
7. PETRI M, ORBAI AM, ALARCÓN GS *et al*.: Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. *Arthritis Rheum* 2012; 64(8): 2677-86. https://doi.org/10.1002/art.34473
8. ARINGER M, COSTENBADER K, DAIKH D *et al*.: 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Ann Rheum Dis* 2019; 78(9): 1151-59. https://doi.org/10.1136/annrheumdis-2018-214819
9. HOCHBERG MC: Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1997; 40(9): 1725. https://doi.org/10.1002/art.1780400928
10. ARINGER M, COSTENBADER K, DAIKH D *et al*.: 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for systemic lupus erythematosus. *Arthritis Rheumatol* 2019; 71(9): 1400-12. https://doi.org/10.1002/art.40930
11. NGIAM KY, KHOR IW: Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019; 20(5): e262-e273. https://doi.org/10.1016/s1470-2045(19)30149-4
12. LING H, GUO ZY, TAN LL, GUAN RC, CHEN JB, SONG CL: Machine learning in diagnosis of coronary artery disease. *Chin Med J* (Engl) 2020; 134(4): 401-3. https://doi.org/10.1097/cm9.0000000000001202
13. ARRIAGA-PIZANO LA, GONZALEZ-OLVERA MA, FERAT-OSORIO EA *et al*.: Accurate diagnosis of sepsis using a neural network: Pilot study using routine clinical variables. *Comput Methods Programs Biomed* 2021; 210: 106366. https://doi.org/10.1016/j.cmpb.2021.106366
14. ZHU L, ZHANG L, HU W *et al*.: A multi-task two-path deep learning system for predicting the invasiveness of craniopharyngioma. *Comput Methods Programs Biomed* 2022, 216:106651. https://doi.org/10.1016/j.cmpb.2022.106651
15. VAN WYK F, KHOJANDI A, KAMALESWARAN R: Improving prediction performance using hierarchical analysis of real-time data: a sepsis case study. *IEEE J Biomed Health Inform* 2019; 23(3): 978-86. https://doi.org/10.1109/jbhi.2019.2894570
16. PHILLIPS M, MARSDEN H, JAFFE W *et al*.: Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open* 2019; 2(10): e1913436. https://doi.org/10.1001/jamanetworkopen.2019.13436
17. CECCARELLI F, LAPUCCI M, OLIVIERI G *et al*.: Can machine learning models support physicians in systemic lupus erythematosus diagnosis? Results from a monocentric cohort. *Joint Bone Spine* 2022; 89(3): 105292. https://doi.org/10.1016/j.jbspin.2021.105292
18. MA Y, CHEN J, WANG T *et al*.: Accurate machine learning model to diagnose chronic autoimmune diseases utilizing information from B cells and monocytes. *Front Immunol* 2022; 13: 870531. https://doi.org/10.3389/fimmu.2022.870531
19. CAI G, ZHU Y, WU Y, JIANG X, YE J, YANG D: A multimodal transformer to fuse images and metadata for skin disease classification. *Vis Comput* 2022; 1-13. https://doi.org/10.1007/s00371-022-02492-4
20. LAO X, MA L, MA Q *et al*.: Hematological factors associated with immunity, inflammation, and metabolism in patients with systemic lupus erythematosus: data from a Zhuang cohort in Southwest China. *J Clin Lab Anal* 2020; 34(6): e23211. https://doi.org/10.1002/jcla.23211
21. YANG Z, LIANG Y, LI C, XI W, ZHONG R: Associations of serum urea, creatinine and uric acid with clinical and laboratory features in patients with systemic lupus erythematosus. *Rheumatol Int* 2012; 32(9): 2715-23. https://doi.org/10.1007/s00296-011-1987-7
22. ARATHI N, SASIDHARAN PK, GEETHA P: Kozhikode criteria for diagnosing systemic lupus erythematosus as a hematological disorder. *J Blood Med* 2016; 7: 13-18. https://doi.org/10.2147/jbm.s95839
23. ALMLÖF JC, ALEXSSON A, IMGENBERG-KREUZ J *et al*.: Novel risk genes for systemic lupus erythematosus predicted by random forest classification. *Sci Rep* 2017; 7(1): 6236. https://doi.org/10.1038/s41598-017-06516-1

24. CECCARELLI F, SCIANDRONE M, PERRICONE C *et al*.: Prediction of chronic damage in systemic lupus erythematosus by using machine-learning models. *PLoS One* 2017; 12(3): e0174200. https://doi.org/10.1371/journal.pone.0174200
25. MAFFI M, TANI C, CASCARANO G *et al*.: Which extra-renal flare is "difficult to treat" in systemic lupus erythematosus? A one-year longitudinal study comparing traditional and machine learning approaches. *Rheumatology* (Oxford) 2023 Apr 24. https://doi.org/10.1093/rheumatology/kead166
26. WILAND P: [Musculoskeletal symptoms in systemic lupus erythematosus and their differential diagnosis with rheumatoid arthritis]. *Ann Acad Med Stetin* 2010; 56 Suppl. 1: 40-44.
27. MESA A, FERNANDEZ M, WU W, NARASIMHAN G, GREIDINGER EL, MILLS DK: Can SLE classification rules be effectively applied to diagnose unclear SLE cases? *Lupus* 2017; 26(2): 150-62. https://doi.org/10.1177/0961203316655212
28. MATTHIESEN R, LAUBER C, SAMPAIO JL *et al*.: Shotgun mass spectrometry-based lipid profiling identifies and distinguishes between chronic inflammatory diseases. *EBioMedicine* 2021; 70: 103504. https://doi.org/10.1016/j.ebiom.2021.103504
29. HUANG Z, SHI Y, CAI B *et al*.: MALDI-TOF MS combined with magnetic beads for detecting serum protein biomarkers and establishment of boosting decision tree model for diagnosis of systemic lupus erythematosus. *Rheumatology* (Oxford) 2009; 48(6): 626-31. https://doi.org/10.1093/rheumatology/kep058
30. DAI Y, HU C, WANG L *et al*.: Serum peptidome patterns of human systemic lupus erythematosus based on magnetic bead separation and MALDI-TOF mass spectrometry analysis. *Scand J Rheumatol* 2010; 39(3): 240-46. https://doi.org/10.3109/03009740903456292
31. ADAMICHOU C, GENITSARIDI I, NIKOLOPOULOS D *et al*.: Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. *Ann Rheum Dis* 2021; 80(6): 758-66. https://doi.org/10.1136/annrheumdis-2020-219069
32. GONG H, SHI C, ZHOU Z *et al*.: Evaluating hypercoagulability in new-onset systemic lupus erythematosus patients using thromboelastography. *J Clin Lab Anal* 2020; 34(5): e23157. https://doi.org/10.1002/jcla.23157
33. MELKI I, ALLAEYS I, TESSANDIER N *et al*.: FcγRIIA expression accelerates nephritis and increases platelet activation in systemic lupus erythematosus. *Blood* 2020; 136(25): 2933-45. https://doi.org/10.1182/blood.2020004974
34. LE MINH G, PESHKOVA AD, ANDRIANOVA IA *et al*.: Impaired contraction of blood clots as a novel prothrombotic mechanism in systemic lupus erythematosus. *Clin Sci* (Lond) 2018; 132(2): 243-54. https://doi.org/10.1042/cs20171510
35. MAO YM, SHI PL, WU L *et al*.: Prevalence and influential factors of thrombocytopaenia in systemic lupus erythematosus patients: a retrospective analysis of 3140 cases in a Chinese population. *Lupus* 2020; 29(7): 743-50. https://doi.org/10.1177/0961203320922301
36. RAMIREZ GA, CANTI V, DEL ROSSO S *et al*.: Diagnostic performance of aPS/PT antibodies in neuropsychiatric lupus and cardiovascular complications of systemic lupus erythematosus. *Autoimmunity* 2020; 53(1): 21-27. https://doi.org/10.1080/08916934.2019.1696778
37. FUJIWARA K, SHIMIZU J, TSUKAHARA H, SHIMADA A: Lupus anticoagulant-hypoprothrombinemia syndrome and immunoglobulin-A vasculitis: a report of Japanese sibling cases and review of the literature. *Rheumatol Int* 2019; 39(10): 1811-19. https://doi.org/10.1007/s00296-019-04404-7
38. RAMOS-CASALS M, BRITO-ZERÓN P, KOSTOV B *et al*.: Google-driven search for big data in autoimmune geoepidemiology: analysis of 394,827 patients with systemic autoimmune diseases. *Autoimmun Rev* 2015; 14(8): 670-79. https://doi.org/10.1016/j.autrev.2015.03.008
39. LECH M, ANDERS HJ: The pathogenesis of lupus nephritis. *J Am Soc Nephrol* 2013; 24(9): 1357-66. https://doi.org/10.1681/asn.2013010026
40. WENDERFER SE, ELDIN KW: Lupus nephritis. *Pediatr Clin North Am* 2019; 66(1): 87-99. https://doi.org/10.1016/j.pcl.2018.08.007
41. KLEIN A, MOLAD Y: Hematological manifestations among patients with rheumatic diseases. *Acta Haematol* 2021; 144(4): 403-12. https://doi.org/10.1159/000511759
42. CASCIO MJ, DeLOUGHERY TG: Anemia: evaluation and diagnostic tests. *Med Clin North Am* 2017; 101(2): 263-84. https://doi.org/10.1016/j.mcna.2016.09.003
43. VELO-GARCÍA A, CASTRO SG, ISENBERG DA: The diagnosis and management of the haematologic manifestations of lupus. *J Autoimmun* 2016; 74: 139-60. https://doi.org/10.1016/j.jaut.2016.07.001
44. CAPSONI F, SARZI-PUTTINI P, ZANELLA A: Primary and secondary autoimmune neutropenia. *Arthritis Res Ther* 2005; 7(5): 208-14. https://doi.org/10.1186/ar1803
45. ARINGER M: Inflammatory markers in systemic lupus erythematosus. *J Autoimmun* 2020; 110: 102374. https://doi.org/10.1016/j.jaut.2019.102374
46. CHEN Y, WANG L, CAO Y, LI N: Total glucosides of paeonia lactiflora for safely reducing disease activity in systemic lupus erythematosus: a systematic review and meta-analysis. *Front Pharmacol* 2022; 13: 834947. https://doi.org/10.3389/fphar.2022.834947