Application of logistic regression and machine learning methods for idiopathic inflammatory myopathies malignancy prediction

W. Zhang[1], G. Huang[2], K. Zheng[1], J. Lin[1], S. Hu[1], S. Zheng[1], G. Du[3], G. Zhang[4], C. Bruni[5], M. Matucci-Cerinic[5,6], D.E. Furst[5,7,8], Y. Wang[1]

[1]Department of Rheumatology and Immunology, Shantou Central Hospital, Shantou, Guangdong, China; [2]Department of Blood Purification, Shantou Central Hospital, Shantou, Guangdong, China; [3]Department of Radiology, Shantou Central Hospital, Shantou, Guangdong, China; [4]Department of Pathology, Shantou University Medical College, Shantou, Guangdong, China; [5]Department of Experimental and Clinical Medicine, Division of Rheumatology, University of Florence, Italy; [6]Unit of Immunology, Rheumatology, Allergy and Rare diseases (UnIRAR), IRCCS San Raffaele Hospital, Milan, Italy; [7]Division of Rheumatology, Department of Medicine, University of California at Los Angeles, CA, USA; [8]University of Washington, Seattle, WA, USA.

Abstract

Objective

Malignancy is related to idiopathic inflammatory myopathies (IIM) and leads to a poor prognosis. Early prediction of malignancy is thought to improve the prognosis. However, predictive models have rarely been reported in IIM. Herein, we aimed to establish and use a machine learning (ML) algorithm to predict the possible risk factors for malignancy in IIM patients.

Methods

We retrospectively reviewed the medical records of 168 patients diagnosed with IIM in Shantou Central hospital, from 2013 to 2021. We randomly divided patients into two groups, the training sets (70%) for construction of the prediction model, and the validation sets (30%) for evaluation of model performance. We constructed six types of ML algorithms models and the AUC of ROC curves were used to describe the efficacy of the model. Finally, we set up a web version using the best prediction model to make it more generally available.

Results

According to the multi-variable regression analysis, three predictors were found to be the risk factors to establish the prediction model, including age, ALT<80U/L, and anti-TIF1-γ, and ILD was found to be a protective factor. Compared with five other ML algorithms models, the traditional algorithm logistic regression (LR) model was as good or better than the other models to predict malignancy in IIM. The AUC of the ROC using LR was 0.900 in the training set and 0.784 in the validation set. We selected the LR model as the final prediction model. Accordingly, a nomogram was constructed using the above four factors. A web version was built and can be visited on the website or acquired by scanning the QR code.

Conclusion

The LR algorithm appears to be a good predictor of malignancy and may help clinicians screen, evaluate and follow up high-risk patients with IIM.

Key words

machine learning, malignancy, idiopathic inflammatory myopathies

*Weijin Zhang, MD**
*Guohai Huang, MD**
Kedi Zheng, MD
Jianqun Lin, MD
Shijian Hu, MD
Shaoyu Zheng, MD
Guangzhou Du, MD
Guohong Zhang, MD
Cosimo Bruni, MD, PhD
Marco Matucci-Cerinic, MD, PhD
Daniel E. Furst, MD
Yukai Wang, MD

**These authors contributed equally and are to be considered co-first authors.*

Please address correspondence to:
Yukai Wang
Department of Rheumatology and Immunology,
Shantou Central Hospital,
no. 114 Waima Road,
515041 Shantou,
Guangdong, China.
E-mail: stzxyywyk@126.com

Received on January 26, 2023; accepted in revised form on February 13, 2023.

Competing interests: C. Bruni reports consultancy fees from Boehringer-Ingelheim and Eli Lilly; grants from Gruppo Italiano Lotta alla Sclerodermia (GILS), Fondazione Italiana Ricerca sull'Artrite (FIRA), European Scleroderma Trial and Research (EUSTAR), Foundation for Research in Rheumatology (FOREUM), Italian Society for Rheumatology (SIR), Scleroderma Clinical Trials Consortium (SCTC), Scleroderma Research Foundation (SRF) outside the submitted work.
M. Matucci-Cerinic has received honoraria from Eli Lilly, BI and Galapagos.
D.E. Furst has received research support from Actelion, Amgen, BMS, Prometheus, Galapagos, GSK, NIH, Novartis, Pfizer, Sanofi, Roche/Genentech, Horizon and Emerald; he has received consultancies from Actelion, Amgen, BMS, Corbus, Galapagos, Novartis, Pfizer and Horizon, and has been member of speaker's bureau for CME only.
The other authors have declared no competing interests.

Introduction

The idiopathic inflammatory myopathies (IIM) are a heterogeneous group of autoimmune rheumatic diseases involving skeletal muscle, the respiratory system, skin and joints (1-3). The subtypes of these IIM patients have different clinical characteristics, including proximal muscle weakness, rapid progressive interstitial lung disease and severe skin lesions. Of note, one distinguishing feature of IIM, particularly dermatomyositis (DM) or polymyositis (PM), is a significant association with a risk of cancer (4). The relationship between DM and cancer in a patient with skin rash, muscle weakness and gastric carcinoma was first reported by Stertz in 1916 (5). Since then, a variety of malignancies have been reported to be closely related with IIM, involving multiple systems, including the nasopharyngeal, lung, breast and gastrointestinal systems (6). In a meta-analysis of case control and cohort studies including 4538 IIM patients in 5 studies, the overall standardised incidence ratio (SIR) as a risk for cancer was 4.66 and 1.75 for DM and PM correspondingly (7). Moreover, the incidence of cancer is highest in the first year after IIM diagnosis (8), and its prognosis is extremely poor owing to the complexity of these two diseases and the discrepancy between tumour and IIM treatment. Hence, it is of great importance to predict the risk of malignance in IIM patients as early as possible.

Recently, various studies have demonstrated the close relationship between cancer and risk factors with regard to multiple demographics, clinical and laboratory features. Patients with IIM onset after 50 years old and male gender may be at higher risk for developing cancer (9, 10). In addition, increased risk of malignancy is associated with skin involvement, with skin necrosis as the strongest association (9). Higher levels of inflammatory markers such as C-reactive protein, erythrocyte sedimentation rate, and creatine kinases were also often observed in IIM patients with malignancy (11, 12). Numerous myositis-associated antibodies have been discovered and verified to indicate different phenotypes of IIM, with some indicators strongly suggesting a high risk of cancer. Anti-p155/140 (anti-TIF1-γ) is associated with the highest positive rate in patients with cancer-associated myositis and this is the predominant diagnostic serological indicator for malignancy (13). This relationship between TIF1-γ and cancer-associated myositis is so tight with odds ratios reaching as high as 23 (95% CI 5.23-101.2) (14). Recent studies also showed a higher prevalence of malignancy in IIM patients with anti-nuclear matrix proteins (NXP)-2 and anti-3-hydroxy-3-methyglutaryl-coenzyme A reductase (HMGCR) antibodies. To date, a quantitative predictive model has rarely been developed to predict the risk for malignancy in IIM patients. A nomogram risk prediction model by Zhong *et al*. (15) showed that patients older than 50-year-old, dysphagia, refractory itching and elevated creatine kinase were risk factors, while interstitial lung disease was a protective factor for dermatomyositis-related-malignancy, with an area under curve (AUC) of 0.756. However, this model did not incorporate TIF1-γ and only applied to patients with DM.

Notably, machine learning (ML) algorithms have been widely utilised in recent years to develop predictive models which appear to have better predictive ability than the traditional regression approaches (16). In this retrospective, case-control study, we aimed to use machine learning to predict and compare algorithms to establish the best risk factors algorithm for malignancy in IIM patients.

Materials and methods

Patients

We retrospectively reviewed the medical records of patients diagnosed with IIM in Shantou Central Hospital, China, from 2013 to 2021. We included 168 patients after excluding 1 patient with too much missing information. This study was approved by the Shantou Central Hospital Ethics Committee (no. 2022-037). Patients included into this study met the classification criteria for IIM (1), including dermatomyositis (DM), polymyositis (PM), immune-mediated necrotising myopa-

thy (IMNM), anti-synthetase syndrome (ASS), and inclusion body myositis (IBD). Exclusion criteria were: 1) absence of complete clinical data; 2) unconfirmed diagnosis of IIM; 3) diagnosis of hepatitis.

Data collection

The following data were collected: (i) baseline information including age and gender; (ii) clinical symptoms involving muscle weakness, myalgia, arthralgia, rash (typical skin involvement of Gottron's rash, Gottron's sign, mechanical hand, heliotrope rash, V-neck sign, shawl sign and holster sign), pruritus, dry mouth and dry eye, dysphagia, respiratory syndrome, fever, oedema, cutaneous ulcer, and Raynaud phenomenon; (iii) clinical signs including rash. For high-risk IIM patients, especially those with several risk factors including the elderly, DM, dysphagia, tumour markers positivity and TIF-1γ positivity, patients were required to be screened for malignancy through PET/CT, as well as gastrointestinal endoscope if digestive symptoms occurred or with their consent. Otherwise, especially those with protective factors including interstitial lung disease (ILD) and negative TIF-1γ antibody, age-appropriate screening, including nasopharyngeal MR, chest CT, abdominal CT, gastrointestinal endoscope, breast ultrasound and thyroid ultrasound were performed as clinically indicated; (iv) laboratory data including white blood cell (WBC), lymphocyte (LY), alanine transaminase (ALT), aspartate aminotransferase (AST), creatinine (Cr), blood urea nitrogen (BUN), lactic dehydrogenase (LDH), creatine kinase (CK), D-dimer, C-reaction protein (CRP), erythrocyte sedimentation rate (ESR), ferritin, carcino-embryonic antigen (CEA), alpha fetoprotein (AFP), carbohydrate antigen 199 (CA199), carbohydrate antigen 125 (CA125), complement 3 (C3), complement 4 (C4), antinuclear antibody (ANA) and myositis antibody profile.

Statistical analysis

Statistical analysis was performed using R (v. 4.05) and SPSS 22.0 (IBM, USA) software. Continuous variables were expressed as mean±SD and were analysed by Student's t-test when they were normally distributed. Variables were analysed by nonparametric methods and described using medians (Q1, Q3) when their distribution was skewed or kurtotic. Categorical variables were analysed using χ^2. When univariate analysis revealed variables with a $p<0.05$, they were included in the predictive models and $p>0.1$ as the criterion for removing variables.

Odds ratios (ORs) and 95% confidence intervals (95% CIs) were calculated for all potential predictors of malignancy when using multivariate regression. When the regression showed a variable to have a $p<0.10$, it was included in the prediction model.

Table I. Demographics, subtypes and malignancy distribution in IIM patients.

	n =168
Age (years, IQR)	56.0 (44.0, 64.8)
Male n (%)	50 (29.8)
Diagnosis	
DM n (%)	86 (51.2)
PM n (%)	40 (23.8)
ASS n (%)	29 (17.2)
IMNM n (%)	13 (7.7)
Malignancy (n=37)	
Nasopharyngeal cancer n (%)	11 (29.7)
Breast cancer n (%)	10 (27.0)
Lung cancer n (%)	7 (18.9)
Oesophagus cancer n (%)	5 (13.5)
Cervical adenocarcinoma n (%)	1 (2.7)
Ovarian cancer n (%)	1 (2.7)
Mediastinum cancer n (%)	1 (2.7)
Multiple cancers n (%)	1 (2.7)
Time relationship between tumourigenesis and disease diagnosis	
simultaneous n (%)	20 (54.1)
before n (%)	5 (13.5)
after n (%)	12 (32.4)

DM: dermatomyositis, PM: polymyositis, ASS: anti-synthetase syndrome, IMNM: immune-mediated necrotic myopathy.

The R packages of "glmnet", "rms", "caret", "rpart", "partykit", "e1071", "MASS", "randomForest", "xgboost", and "neuralnet" were used to establish the prediction model of ML algorithms. The R packages of "pROC" and "rmda" were used to validate the prediction ability of the model. The R packages of "corrplot" and "ggcorplot" were used to establish the heat map. The R package of "shiny" and "shinyPredict" was used to establish the web application. The R packages of "ingredients" and "DALEX" were used to show the relative importance of variables of prediction model.

In this study, we randomly split patients into two groups, namely the training sets (70%) for construction of prediction model, and the validation sets (30%) for evaluation of model performance. We constructed six types of ML algorithms models, *i.e.* Logistic regression (LR), Support vector machine (SVM), random forest (RF), Classification and regression tree (CART), Extreme gradient boosting (XGBoost), and Neural network (NNET). Then we used the area under curve (AUC) of the receiver operating characteristic (ROC) curve to evaluate and compare the predictive ability of the models in the training and validation sets. The value of the AUC of the ROC curve was used to describe the efficacy of the model. Finally, we set up the web version using the best prediction model. The prediction probability of malignancy in IIM can be easily calculated and displayed on the website after inputting clinical features.

Results

Population characteristics

We identified 168 patients with IIM diagnosed between 2013 and 2021. Twen-

Table II. Prevalence rate of malignancy in different subtypes of IIM patients.

Category	DM	PM	ASS	IMNM	*p*
n	86	40	29	13	
male n (%)	24 (27.9)	17 (42.5)	4 (13.8)	5 (38.5)	0.055
age (year)	56.0 (44.0, 63.0)	57.0 (41.3, 63.8)	58.0 (49.5, 66.5)	41.0 (31.5, 61.5)	0.137
malignancy n (%)	27 (31.4)	4 (10.0)	6(20.7)	0	0.002*

DM: dermatomyositis, PM: polymyositis, ASS: anti-synthetase syndrome, IMNM: immune-mediated necrotic myopathy. *$p<0.05$.

Table III. Patients' characteristics and symptoms in the malignancy group and non-malignancy group. Data are expressed with interquartile range (Q1, Q3) if the distribution was abnormal, and otherwise with mean ± SD for continuous data. For categorical variables, data are expressed with number (%).

	Malignancy n=37	Non-malignancy n=131	Statistics	*p*
Age (years, IQR)	59.0 (51.5, 65.0)	55.0 (39.0, 63.0)	2.100	0.036*
Male n (%)	15 (40.5)	35 (26.7)	2.637	0.104
DM n (%)	27 (73.0)	59 (45.0)	9.011	0.003*
Myasthenia n (%)	28 (75.7)	87 (66.4)	1.146	0.284
Myalgia n (%)	15 (40.5)	64 (48.9)	0.801	0.371
Arthralgia n (%)	2 (5.4)	32 (24.4)	6.467	0.011*
Pruritus n (%)	10 (27.0)	26 (19.8)	0.883	0.347
Dry mouth and dry eye n (%)	3 (8.1)	12 (9.2)	0.039	0.843
Dysphagia n (%)	13 (35.1)	23 (17.6)	5.295	0.021*
Respiratory involvement n (%)	7 (24.3)	54 (41.2)	3.515	0.040*
Fever n (%)	0	11 (8.4)	2.094	0.148
Oedema n (%)	3 (8.1)	4 (3.1)	0.797	0.372
Cutaneous ulcer n (%)	2 (5.4)	5 (3.8)	0.000	1.000
Raynaud phenomenon n (%)	0	6 (4.6)	0.679	0.410
Gottron's rash n (%)	13 (35.1)	33 (25.2)	1.435	0.231
Gottron's sign n (%)	18 (48.6)	35 (26.7)	6.426	0.011*
Mechanical hand n (%)	4 (10.8)	16 (12.2)	0.000	1.000
Heliotrope rash n (%)	18 (48.6)	42 (32.1)	3.458	0.063
V-neck sign n (%)	18 (48.6)	23 (17.6)	15.117	<0.001*
Shawl sign n (%)	11 (29.7)	17 (13.0)	5.830	0.016*
Holster sign n (%)	3 (8.1)	6 (4.6)	0.708	0.400
ILD n (%)	11 (29.7)	75 (57.3)	8.747	0.003*

DM: dermatomyositis; ILD: interstitial lung disease. *p<0.05.

Table IV. Comparison of laboratory data between malignancy group and non-malignancy group.

	Malignancy n=37	Non-malignancy n=131	Statistics	*p*
WBC 10^9/L	7.68 ± 3.14	8.71 ± 4.27	-1.369	0.173
LY 10^9/L	1.2 (0.6, 1.7)	1.3 (0.9, 2.0)	-1.188	0.235
ALB g/L n=166	35.03 ± 5.75	35.2 ± 6.4	-0.102	0.919
ALT U/L	35.0 (19.0, 69.0)	68.0 (29.0, 160.0)	-2.540	0.011*
AST U/L	62.0 (27.0, 128.5)	88.0 (37.0, 190.0)	-1.678	0.093
Cr umol/L	50.7 (45.6, 70.4)	55.0 (42.9, 66.7)	0.239	0.811
BUN mmol/L	4.3 (3.6, 5.5)	4.5 (3.5, 6.0)	-0.136	0.892
LDH U/L n=163	405.0 (318.0, 705.5)	459.0 (331.0, 777.0)	-0.750	0.453
CK U/L n=165	724.0 (156.3, 2489.0)	696.0 (171.0, 4830.0)	-0.572	0.567
D-dimer ug/L n=138	880.0 (430.0, 2020.0)	696.0 (432.5, 1662.5)	0.207	0.836
CRP mg/L n=159	5.5 (3.0, 12.3)	6.4 (2.3, 14.6)	0.465	0.642
ESR mm/h n=147	14.0 (9.0, 29.0)	22.0 (10.0, 45.0)	-1.928	0.054
Ferritin ng/mL n=110	704.4 (382.0, 1033.3)	583.3 (279.5, 1217.0)	0.960	0.337
CEA ng/mL n=161	2.4 (1.6, 4.5)	1.8 (1.2, 3.1)	1.889	0.059
AFP IU/ml n=160	2.1 (1.5, 3.0)	1.9 (1.4, 3.2)	0.713	0.476
CA199 U/mL n=131	7.5 (5.5, 16.2)	9.3 (5.4, 18.4)	-0.628	0.530
CA125 U/mL n=130	10.2 (7.2, 15.7)	11.2 (7.3, 18.6)	-0.691	0.489
C3 g/L n=146	0.91 ± 0.17	0.88 ± 0.22	0.667	0.661
C4 g/L n=146	0.23 ± 0.07	0.22 ± 0.08	0.439	0.914

WBC: white blood cell; LY: lymphocyte; ALB: albumin; ALT: alanine transaminase; AST: aspartate aminotransferase; Cr: creatinine; BUN: blood urea nitrogen; LDH: lactic dehydrogenase; CK: creatine kinase; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; CEA: carcino-embryonic antigen; AFP: alpha fetoprotein; CA199: carbohydrate antigen 199; CA125: carbohydrate antigen 125; C3: complement 3; C4: complement 4. *p<0.05.

ty-nine and eight tenths of them are male. Median age at IIM diagnosis was 56.0 (44.0, 64.8) (Table I). Eighty-six patients (51.2%) were classified as DM (with 31.4% malignancy), 40 patients (23.8%) as PM (with 10% malignancy) and 29 patients (17.2%) as ASS (with 20.7% malignancy) the rest (7.7%) as other IIM patients (Table II). Among these 168 patients, 37 patients had a malignancy (the Malignancy group). The top three malignant tumours were nasopharyngeal cancer (29.7%), breast cancer (27%) and lung cancer (18.9%). The remaining 131 patients were designated as the Non-Malignancy group. When contrasting these two groups, we found that patients with malignancy were statistically significantly older and more frequently diagnosed as DM (p<0.05 for both). Clinically, patients with the following characteristics were more likely to develop malignancies: dysphagia (p=0.021), Gottron's sign (p=0.011), V-neck sign (p<0.001), and shawl sign (p=0.016). In contrast, the following characteristics were associated with a lower likelihood of malignancy: arthralgia (p=0.011), respiratory involvement (p=0.04), and ILD (p=0.003). There were no differences in gender, muscle weakness, myasthenia, myalgia, pruritus, dry mouth and dry eye, fever, oedema, cutaneous ulcer, Raynaud's phenomenon, mechanical hand, heliotrope rash, and holster sign (Table III).

Among the laboratory data, the Malignancy group had statistically higher likelihood of a positive TIF1-γ (p<0.001) and a lower ALT (p=0.011). Categorical variables of ALT level were adopted because it did not meet with linear correlation, and ALT<80U/L was selected as the threshold. The two groups did not differ in term of WBC, LY, AST, Cr, BUN, LDH, CK, D-dimer, CRP, ESR, ferritin, CEA, AFP, CA199, CA125, C3, C4 and the rest of the myositis antibody profile (Table IV and V).

Risk factors for malignancy among IIM patients

In the univariable analysis, age, DM, arthralgia, dysphagia, respiratory involvement, Gottron's sign, V-neck sign, shawl sign, ILD, ALT<80U/L, CEA>2.0 ng/ml and anti-TIF1-γ were statistically significantly different and were included in the multi-variable regression analysis (Table VI). The relationships among variables are illustrated by a heat map analysis (Fig. 1). Based on the multivariable analysis, the only independent risk factors that

Table V. Comparison of myositis antibody profile between malignancy group and non-malignancy group.

	Malignancy n=37	Non-malignancy n=131	Statistics	*p*
ANA (n=162)	21 (60.0)	69 (54.3)	0.357	0.550
MDA5	1 (3.8)	19 (17.9)	2.217	0.137
TIF1γ	10 (38.5)	4 (3.8)	22.965	<0.001*
NXP2	0	9 (8.5)	1.221	0.269
SAE1	1 (3.8)	1 (0.9)		0.356
Mi-2	0	4 (3.8)		0.585
ARS				
EJ	0	4 (3.8)		0.585
OJ	1 (3.8)	1 (0.9)		0.356
PL-7	2 (7.7)	4 (3.8)	0.112	0.738
PL-12	0	1 (0.9)		1.000
Jo-1 (n=161)	4 (11.4)	23 (18.3)	0.914	0.339
HA	0	1 (0.9)		1.000
SRP	0	11 (10.4)	1.767	0.121
HMGCR	0	1 (0.9)		1.000
Cn1a	0	2 (1.9)		1.000
PMSCL75	0	2 (1.9)		1.000
KU	1 (3.8)	1 (0.9)		0.356
RNA-PIII	0	1 (0.9)		1.000
Th/To	0	2 (1.9)		1.000
Ro-52 (n=161)	16 (45.7)	56 (44.4)	0.018	0.894

ANA: antinuclear antibody; MDA5: melanoma differentiation-associated gene 5; TIF1γ: transcription intermediary factor 1γ; NXP2: nuclear matrix protein 2; SAE1: small ubiquitin-like modifier 1; ARS: aminoacyl tRNA synthetases; EJ: glycyl; OJ: isoleucyl; PL-7: threonyl; PL-12: alanyl; Jo-1: histidyl; HA: tyrosyl; SRP: signal recognition particle; HMGCR: 3-hydroxy 3-methylutaryl coenzyme A reductase; Cn1a: cytoplasmic 5' nucleotidase 1A; PMSCL75: polymyositis-scleroderma 75; RNA-PIII: RNA polymerase III. *p<0.05.

Table VI. Univariate and multivariable logistic regression analysis of risk factors for idiopathic inflammatory myopathies. (Step backward, Wald test, entry condition 0.05, deletion condition 0.10).

Factors	Univariate		Multivariate	
	p	OR(95%CI)	*p*	OR (95%CI)
Age (per 10 year)	0.009	1.456 (1.099-1.931)	0.052*	1.612 (0.997-2.607)
DM	0.004	3.295 (1.476-7.355)		
Arthralgia	0.022	0.177 (0.040-0.776)		
Dysphagia	0.024	2.543 (1.130-5.725)		
Respiratory involvement	0.065	0.458 (0.200-1.049)		
Gottron's sign	0.013	2.598 (1.225-5.512)		
V-neck sign	<0.001	4.449 (2.027-9.765)		
Shawl sign	0.019	2.837 (1.189-6.771)		
ILD	0.004	0.316 (0.144-0.693)	0.007*	0.193 (0.058-0.643)
ALT <80U/L	0.006	3.739 (1.460-9.577)	0.024*	11.175 (1.367-91.355)
CEA >2.0 ng/ml	0.031	2.316 (1.081-4.964)		
TIF1γ	<0.001	15.781 (4.415-56.410)	0.028*	4.963 (1.193-20.642)

DM: dermatomyositis; ILD: interstitial lung disease; ALT: alanine transaminase; CEA: carcino-embryonic antigen; TIF1γ: transcription intermediary factor 1γ. *p<0.10.

predicted malignancy were age (per ten years, OR=1.612; 95% CI [0.997, 2.607]; p=0.052- included despite being slightly greater than 0.05 based on the literature (9, 15, 17) and clinical judgement), ALT<80 U/L (OR=11.175; 95% CI [1.367, 91.355]; p=0.024), and anti-TIF1-γ (OR=4.963; 95% CI [1.193, 20.642]; p=0.028). Interstitial lung disease (OR=0.193; 95% CI [0.058, 0.643]; p=0.007) was a negative predictive factor.

The use of machine learning algorithms

The performance of six different ML algorithm models as predictors of malignancy in training sets and validation sets are shown and compared in Figures 2 and 3 and Table VII and VIII, respectively. The results showed that the NNET and RF model possessed excellent predictive ability in the training set, but did not do well in the validation set. The traditional Logistic Regression algorithm model did as well or better than other machine learning algorithms in predicting malignancy of IIM (AUC of ROC was 0.900 in the training set and 0.784 in the validation set) (Table VII and VIII). Therefore, we selected the LR model as the final prediction model.

The relative importance of variables in prediction models

The relative importance of variables in each prediction model is shown in Figure 4. Although the importance of different variables in different models was variable, and differences among the variables were small, anti-TIF1-γ and low ALT were numerically the two most importance positive predictive variables among the 6 models and ILD was most useful as negative predictor.

Prediction and validation of model for malignancy in IIM patients

On the basis of the four factors selected by multivariate analysis (Age, TIF1-γ, ALT and ILD), for the convenience of clinical application, we constructed a nomogram to predict the probability of malignancy in IIM patients (Fig. 5). One determines the numerical value of each factor based on the vertical line intersection between the variable and the point axis, and then adds all variable points to calculate the total risk score, with each risk score corresponding to the probability of malignancy. The usefulness of this nomogram will need to be tested in several other datasets but it seemed useful as applied in our patients.

The web version of model

For extending the application of the model established in this study, a web version was built and can be visited on the website https://hgh-163.shinyapps.io/DynNomapp/ or acquired by scanning the QR code (Fig. 6) with a smartphone. The algorithm deter-

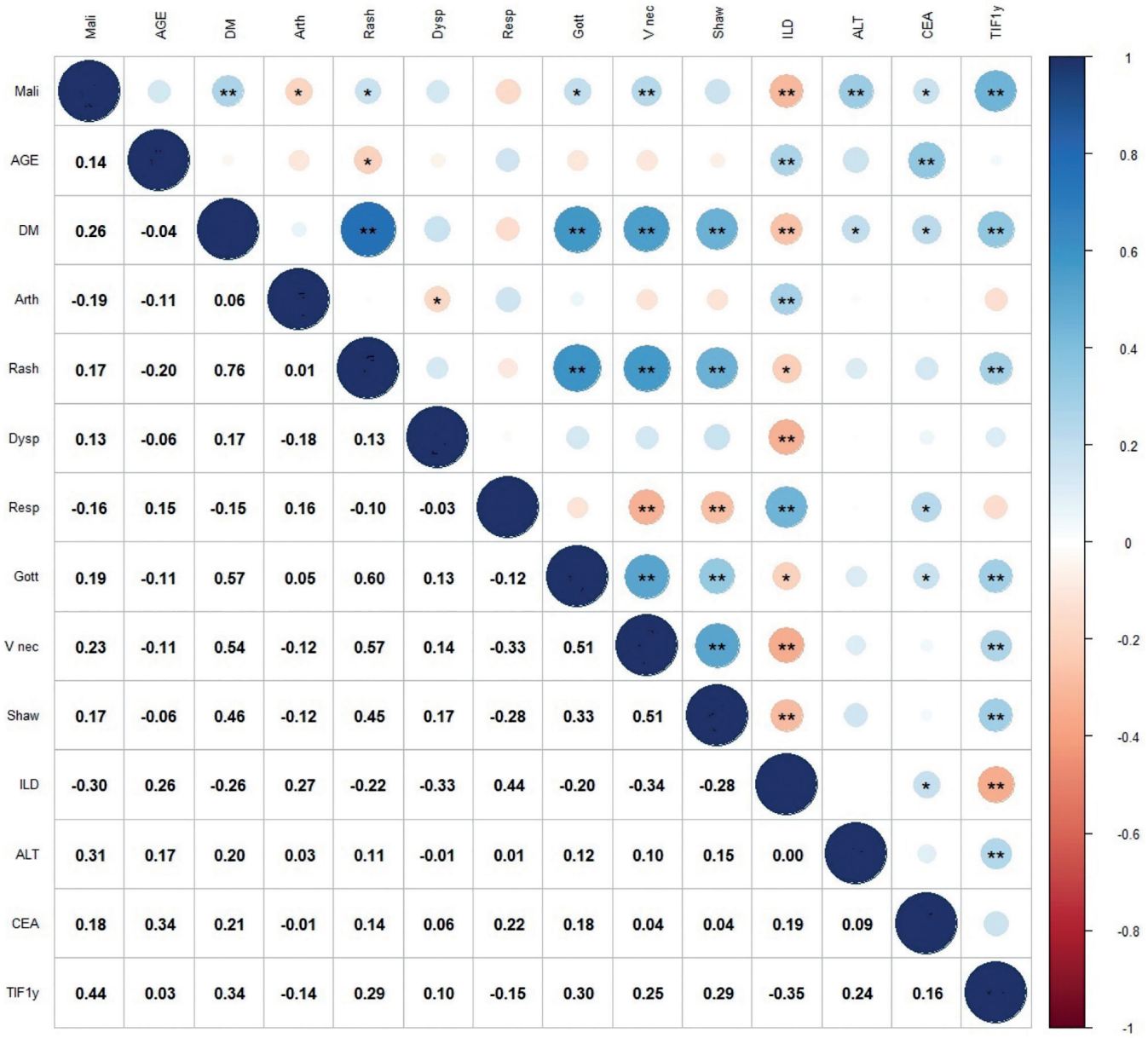

Fig. 1. The relationship between different variables and malignancy.
Mali: malignant; Age: age per ten years; DM: dermatomyositis; Arth: arthralgia; Rash: rash; dysp: dysphagia; Resp: respiratory involvement; Gott: Gottron's sign; V nec: V-neck sign; Shaw: shawl sign; ILD: interstitial lung disease; ALT: alanine transaminase <80U/L; CEA: carcino-embryonic antigen >2.0 ng/ml; TIF1γ: transcription intermediary factor 1γ.
The Pearson coefficient value shown on the left of the heat map. On the right of the heat map, the red ball indicates negative correlation, and the blue ball indicates positive correlation.
The size and colour depth is positively proportional to the correlation coefficient. *p<0.05, **p<0.01.

mines the predicted probability of malignancy by inputting the clinical characteristics.

Discussion

Malignancy may be disguised as inflammatory myositis, and myositis may be a manifestation of the paraneoplastic syndrome of malignancy. Thus, the rashes of IIM may be a skin window for recognising malignancy. About one-third of patients with myositis develop a malignancy within 3 years of "IIM" diagnosis, while patients with IIM have the highest risk of developing malignancy within 1 year of diagnosis (18). Further, malignancy remains one of the leading causes of death in IIM patients (19, 20). Thus, it is of great importance to be able to predict malignancy in IIM patients, and as early as possible, so that treatment of the malignancy can begin quickly, thus increasing the probability of a good outcome.

The aim of our work was to establish, in a retrospective, case-control study, a potentially clinically useful, multivariate risk prediction model for malignancy in IIM patients. After examining and comparing 6 machine learning algorithms, the logistic regression model remained at least as good as and perhaps slightly better than the machine learning algorithms. Based on LR, four factors were selected as the final components in the model to predict malig-

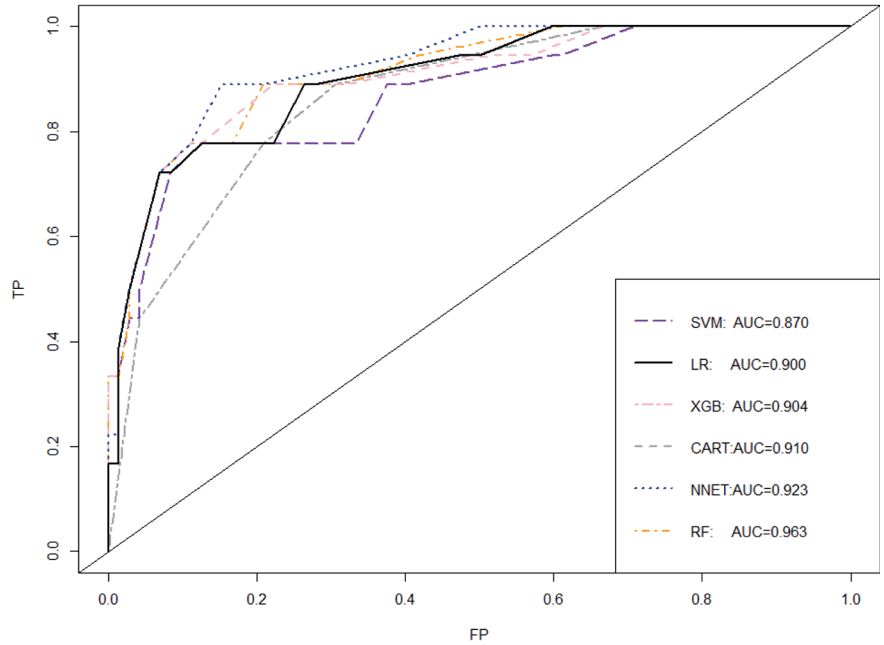

Fig. 2. ROC curve analysis of different machine algorithms prediction model for malignancy of idiopathic inflammatory myopathies in the training set.
SVM: support vector machine; LR: logistic regression; XGBoost: Extreme Gradient Boosting; CART: classification and regression tree; NNET: neural network; RF: random forest.

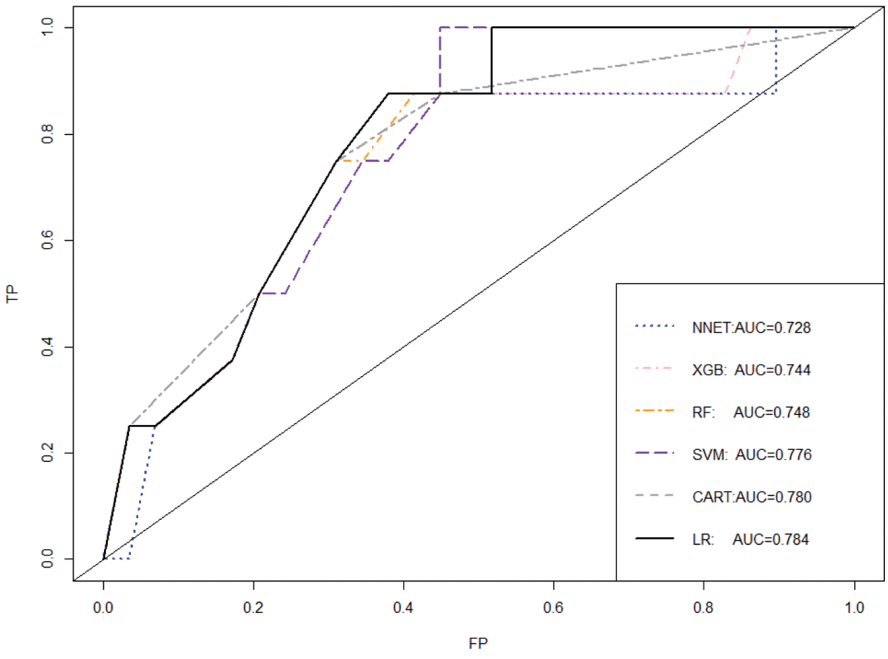

Fig. 3. ROC curve analysis of different machine algorithms prediction model for malignancy of idiopathic inflammatory myopathies in the validation set.
NNET: neural network; XGBoost: Extreme Gradient Boosting; RF: random forest; SVM: support vector machine; CART: classification and regression tree; LR: logistic regression.

nancy in IIM. The presence of ALT <80 U/L, increasing age and anti-TIF1-γ predicted increased probability of malignancy, while the presence of ILD was a protective factor.

Although age was a borderline statistically significant factor in LR, it contributed to the other algorithms and age has been associated with malignancy in the literature. A meta-analysis incorporating 380 IIM patients and 1575 controls in 20 studies showed that older age influences susceptibility to cancer (9). Another study reported that the median age of DM patients with cancer was older than those without cancer (17). In another study, age >50 was selected as one of the factors to construct a nomogram for predicting malignancy in dermatomyositis patients (15). In our study, too, the malignancy group was older than the non-malignancy group, with the median age 59.00 *vs*. 55.00, p=0.036. And, of course, increasing age is associated with malignancies in general. Thus, systemic cancer screening is strongly recommended for IIM patients older than 50 years old.

Most previous studies focused on DM associated malignancy, while we examined the subtypes of IIM- DM, PM, IMNM, ASS and IBM. We found that the prevalence of malignancy in patients with DM was associated with the highest risk of cancer-31.4% in our cohort, which is within the 13% to 42% range found in the literature (7, 21, 22). The V-neck sign, shawl sign and Gottron's sign, which are typical signs of DM, also predicted malignancy (p<0.05). Non-statistical difference in Gottron's rash, holster sign and mechanical hand may be due to small sample size in this study and weaker correlation with malignancy. Thus, a larger and prospective study should be carried out for verification of different rashes in malignancy prediction.

We found a relatively high prevalence of cancer in ASS, with 6 malignancies in 29 patients. This is the second most common prevalence after DM. The literature reports this relationship rarely, perhaps because ASS is a relatively rare disease. The literature reports a prevalence of up to 16.6% and our prevalence was 20.7%, in the same general range as the literature (23).

Malignancy was found in 10.0% of PM patients, ranking the third highest risk, which was also consistent with the reported risk range of 3% to 18% (21, 24). Among the 13 patients with IMNM in our cohort, no cancer was found and IMNM was not a risk factor for malignancy. Of course, the very few patients with IMNM make any predictive algorithm suspect. The lack of malignancy in these patients is supported by the literature, as extramuscular involvement is rare in IMNM. The two serological markers of IMNM,

Table VII. Predictive performance of different machine algorithms model for malignancy of idiopathic inflammatory myopathies in the training set.

Models	Accuracy	Sensitivity	Specificity	AUROC
LR	0.867	0.875	0.866	0.900
CART	0.856	0.727	0.873	0.910
SVM	0.856	0.727	0.873	0.870
RF	0.867	0.800	0.875	0.963
NNET	0.879	0.722	0.918	0.923
XGB	0.878	0.818	0.886	0.904

LR: logistic regression; CART: classification and regression tree; SVM: support vector machine; RF: random forest; NNET: neural network; XGB: Extreme Gradient Boosting.

Table VIII. Predictive performance of different machine algorithms model for malignancy of idiopathic inflammatory myopathies in the validation set.

Models	Accuracy	Sensitivity	Specificity	AUROC
LR	0.784	0.500	0.818	0.784
CART	0.811	0.667	0.824	0.780
SVM	0.811	0.667	0.824	0.776
RF	0.811	0.667	0.824	0.748
NNET	0.730	0.375	0.828	0.728
XGB	0.784	0.500	0.818	0.744

LR: logistic regression; CART: classification and regression tree; SVM: support vector machine; RF: random forest; NNET: neural network; XGB: Extreme Gradient Boosting.

anti-SRP antibody and anti-HMGCR antibody, were not associated with malignancy in a meta-analysis (25) although one author did find such a relationship (26). More research clearly needs to be done in this area.

It was interesting to examine the usefulness of antibodies when predicting malignancies in IIM. The presence of TIF1-γ (155-kDa) increased the risk of a malignancy fivefold, and it was an important variable as a predictor in all models. This finding is supported by the literature. A meta-analysis regarding the usefulness of TIF1-γ in DM showed that 80% of myositis patients with malignancy tested positive for this antibody, and 90% of patients without malignancy tested negative, indicating high sensitivity and specificity (27). Although the sensitivity and specificity in our study were not as high as the above (31.25% and 71.42% respectively), it might be a simple screening tool which would alert the clinician that a given patient needs to be closely followed.

Another well-recognised cancer-associated antibody is NXP-2, which is linked with muscle weakness and elevated CK (28). Ichimura *et al*. (29) collected 445 cases of DM and 62 cases of PM, of whom 7 (1.6%) and 1 (1.6%) tested positive in NXP-2, respectively, although another study found no difference in NXP2 positivity among those with or without cancer (30). Of the 8 patients described by Ichimura *et al*., 3 cases (37.5%) developed visceral malignancies within 3 years of IIM diagnosis, indicating that NXP-2 might be a marker of increased cancer risk in IIM. This hypothesis was supported by Fiorentino *et al*. (31). Thus far, no cancer has been observed in our 9 NXP2 positive patients, with follow-up of up to 8 years. Larger studies are needed to determine the usefulness of NXP2.

In our study, ILD was identified as a protective factor for malignancy in patients with IIM, which was consistent with the medical literature (32, 33). According to a retrospective study by Zhong *et al*., there is a negative correlation between ILD and tumours in DM patients, with an OR value as low as

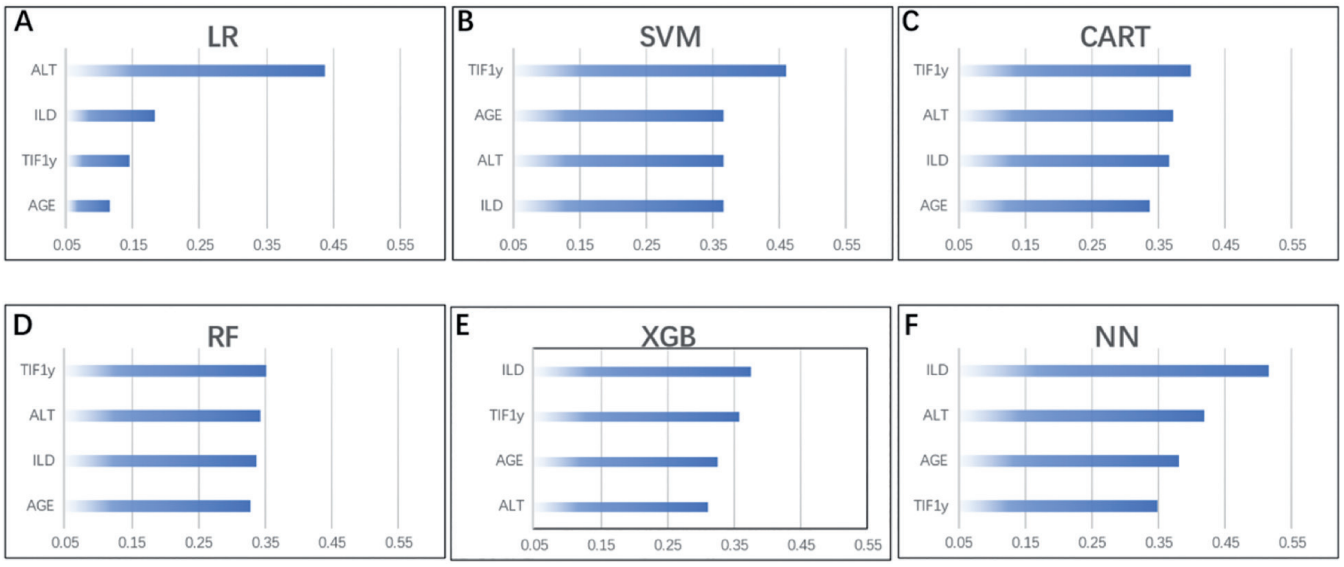

Fig. 4. Relative importance of each variable in different machine algorithms prediction model.
LR: logistic regression (**A**); SVM: support vector machine (**B**); CART: classification and regression tree (**C**); RF: random forest (**D**); XGBoost: Extreme Gradient Boosting (**E**); NN: neural network (**F**).

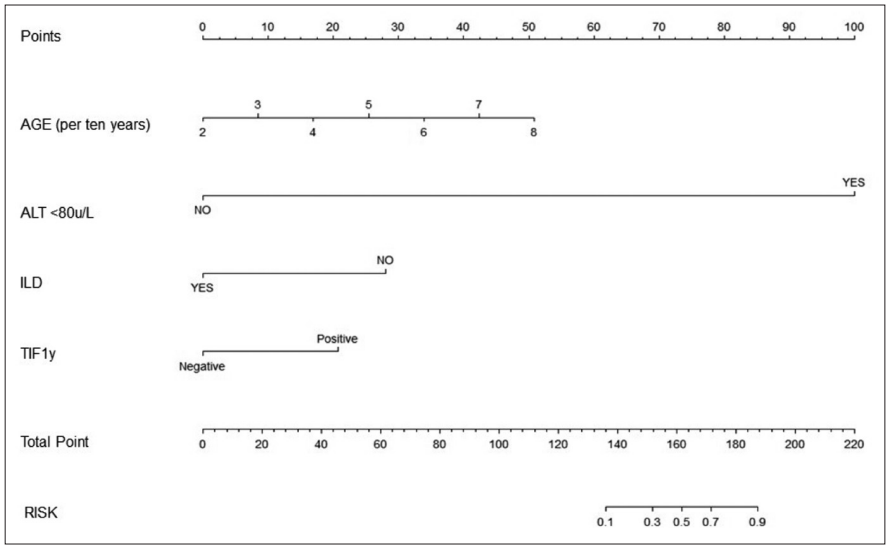

Fig. 5. The nomogram prediction model of the malignancy probability in IIM patients. Determine the value of each factor based on the vertical line intersection between the variable and the point axis, and then add all variable points to calculate the total risk score, with each risk score corresponding to the probability of malignancy.

Fig. 6. QR code to be scanned with a smartphone for nomogram prediction.

0.367 (95% CI [0.147, 0.913]) in multivariable analysis (15). Furthermore, a meta-analysis confirmed the protective role of ILD in cancer of IIM (9). On the other hand, MDA5 positive patients (some with lung disease) have a poor prognosis in DM, and the medications used to treat DM predispose to cancers and ILD per se can increase the risk of lung cancer (34, 35). These conflicting factors are puzzling and certainly will require further research.

Similar to the 151 patient study of So *et al*. (21), our results pointed out that the following were risk factors for malignancy in IIM: (i) older age; (ii) ALT <80 U/L; (iii) the presence of TIF1-γ; and (iv) the absence of ILD. However, we did not find that dysphagia predicted cancer in multivariable analysis. Coincident with our result, their laboratory data also showed that lower serum AST was related with malignancy in IIM patients and this phenomenon was rarely reported. Possibly, IIM patients without malignancy may have more probability to behave with another system involvement, for instance and dominantly in muscle injury wherein the ALT or AST are usually elevated. This variable indeed surprised us due to its high weight among the ML models (21).

Several studies (4, 6, 36) have shown that inflammatory indicators, CRP and ESR, can be used as predictive markers for malignancy. However, it was not the case in our research, perhaps because of the other factors were more important in the multi-variable models, making these not statistically significant. Also, inflammation is not merely associated with the tumour, but also may be related to patient's multiple other concurrent, co-morbidities (*e.g*. infection, necrosis, drugs, other inflammatory conditions) which may confound their usefulness.

Our data have some notable strengths. It includes a relatively robust number of patients which have been sub-set into IIM subtypes and carefully followed and analysed. Also, updated serological testing has been done (*e.g*. TIF1-γ). Further a relatively sophisticated analysis was undertaken using machine learning technology, thus increasing the robustness of the results, increasing the probability of credible results and allowing internal consistency. Further we tried to make the results easily available for clinical use thru the development of the web-based nomogram.

However, our data have some limitations. First, our data comes from a single-centre, which may enrol sicker patients and increase the prevalence of malignancy. Furthermore, because the machine learning algorithm itself is closely related with sample size, better performance of LR was probably due to the relatively small sample. Thus, it needs to be further verified by multicenter studies. Second, this is a retrospective study and data were missing, especially the anti-myositis antibody profiles. Third, the length of follow-up would ideally be longer. Fourth, tests for cryoglobulinaemia and elevated aldolase were not available in our unit and confounding with other diseases may have occurred, although it was unlikely based on clinical results.

This study summarised the possible risk factors for malignancy in a retrospective and case-control study of IIM patients, and established a multivariate risk prediction model of LR, which has good usefulness for clinical application and may help clinicians screen, evaluate and follow up those high-risk patients with IIM. However, it still needs more cases to optimise this model.

References

1. MARIAMPILLAI K, GRANGER B, AMELIN D *et al*.: Development of a new classification system for idiopathic inflammatory myopathies based on clinical manifestations and myositis-specific autoantibodies. *Jama Neurol* 2018; 75: 1528-37. https://doi.org/10.1001/jamaneurol.2018.2598
2. CARDELLI C, ZANFRAMUNDO G, COMETI L *et al*.: Idiopathic inflammatory myopathies: One year in review 2021. *Clin Exp Rheumatol* 2022; 40: 199-209. https://doi.org/10.55563/clinexprheumatol/vskjxi
3. LI Y, LI Y, WANG Y *et al*.: Nomogram to predict dermatomyositis prognosis: A population-based study of 457 cases. *Clin Exp Rheumatol* 2022; 40: 247-53. https://doi.org/10.55563/clinexprheumatol/0ddt88
4. MOGHADAM-KIA S, ODDIS CV, ASCHERMAN DP, AGGARWAL R: Risk factors and cancer screening in myositis. *Rheum Dis Clin North Am* 2020; 46: 565-76.

https://doi.org/10.1016/j.rdc.2020.05.006
5. PRZYBYLSKI G, JARZEMSKA A, CZERNIAK J, SIEMIATKOWSKA K, GADZINSKA A, CIESLINSKI K: A case report of a patient with dermatomyositis as a prodromal sign of lung cancer. *Pol Arch Med Wewn* 2008; 118: 143-7.
6. TINIAKOU E, MAMMEN AL: Idiopathic inflammatory myopathies and malignancy: a comprehensive review. *Clin Rev Allergy Immunol* 2017; 52: 20-33. https://doi.org/10.1007/s12016-015-8511-x
7. QIANG JK, KIM WB, BAIBERGENOVA A, ALHUSAYEN R: Risk of malignancy in dermatomyositis and polymyositis. *J Cutan Med Surg* 2017; 21: 131-6. https://doi.org/10.1177/1203475416665601
8. CHEN D, YUAN S, WU X *et al*.: Incidence and predictive factors for malignancies with dermatomyositis: A cohort from southern China. *Clin Exp Rheumatol* 2014; 32: 615-21.
9. WANG J, GUO G, CHEN G, WU B, LU L, BAO L: Meta-analysis of the association of dermatomyositis and polymyositis with cancer. *Br J Dermatol* 2013; 169: 838-47. https://doi.org/10.1111/bjd.12564
10. ANDRAS C, PONYI A, CONSTANTIN T *et al*.: Dermatomyositis and polymyositis associated with malignancy: A 21-year retrospective study. *J Rheumatol* 2008; 35: 438-44.
11. SPARSA A, LIOZON E, HERRMANN F *et al*.: Routine vs extensive malignancy search for adult dermatomyositis and polymyositis: a study of 40 patients. *Arch Dermatol* 2002; 138: 885-90. https://doi.org/10.1001/archderm.138.7.885
12. GALLAIS V, CRICKX B, BELAICH S: [Prognostic factors and predictive signs of malignancy in adult dermatomyositis]. *Ann Dermatol Venereol* 1996;123:722-6.
13. KAJI K, FUJIMOTO M, HASEGAWA M *et al*.: Identification of a novel autoantibody reactive with 155 and 140 kDa nuclear proteins in patients with dermatomyositis: An association with malignancy. *Rheumatology* (Oxford) 2007; 46: 25-8. https://doi.org/10.1093/rheumatology/kel161
14. TRALLERO-ARAGUAS E, LABRADOR-HORRILLO M, SELVA-O'CALLAGHAN A *et al*.: Cancer-associated myositis and anti-p155 autoantibody in a series of 85 patients with idiopathic inflammatory myopathy. *Medicine* (Baltimore) 2010; 89: 47-52. https://doi.org/10.1097/MD.0b013e3181ca14ff
15. ZHONG J, HE Y, MA J, LU S, WU Y, ZHANG J: Development and validation of a nomogram risk prediction model for malignancy in dermatomyositis patients: A retrospective study. *Peerj* 2021; 9: e12626. https://doi.org/10.7717/peerj.12626
16. WU Y, RAO K, LIU J *et al*.: Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer. *Front Endocrinol* (Lausanne) 2020; 11: 577537. https://doi.org/10.3389/fendo.2020.577537
17. ANTIOCHOS BB, BROWN LA, LI Z, TOSTESON TD, WORTMANN RL, RIGBY WF: Malignancy is associated with dermatomyositis but not polymyositis in Northern New England, USA. *J Rheumatol* 2009; 36: 2704-10. https://doi.org/10.3899/jrheum.090549
18. YANG Z, LIN F, QIN B, LIANG Y, ZHONG R: Polymyositis/dermatomyositis and malignancy risk: A metaanalysis study. *J Rheumatol* 2015; 42: 282-91. https://doi.org/10.3899/jrheum.140566
19. DOBLOUG GC, GAREN T, BRUNBORG C, GRAN JT, MOLBERG O: Survival and cancer risk in an unselected and complete Norwegian idiopathic inflammatory myopathy cohort. *Semin Arthritis Rheum* 2015; 45:3 01-8. https://doi.org/10.1016/j.semarthrit.2015.06.005
20. NUNO-NUNO L, JOVEN BE, CARREIRA PE *et al*.: Mortality and prognostic factors in idiopathic inflammatory myositis: A retrospective analysis of a large multicenter cohort of Spain. *Rheumatol Int* 2017; 37: 1853-61. https://doi.org/10.1007/s00296-017-3799-x
21. SO MW, KOO BS, KIM YG, LEE CK, YOO B: Idiopathic inflammatory myopathy associated with malignancy: A retrospective cohort of 151 Korean patients with dermatomyositis and polymyositis. *J Rheumatol* 2011; 38: 2432-5. https://doi.org/10.3899/jrheum.110320
22. SIGURGEIRSSON B, LINDELOF B, EDHAG O, ALLANDER E: Risk of cancer in patients with dermatomyositis or polymyositis. A population-based study. *N Engl J Med* 1992; 326: 363-7. https://doi.org/10.1056/NEJM199202063260602
23. DUGAR M, COX S, LIMAYE V, BLUMBERGS P, ROBERTS-THOMSON PJ: Clinical heterogeneity and prognostic features of South Australian patients with anti-synthetase autoantibodies. *Intern Med J* 2011; 41: 674-9. https://doi.org/10.1111/j.1445-5994.2010.02164.x
24. STOCKTON D, DOHERTY VR, BREWSTER DH: Risk of cancer in patients with dermatomyositis or polymyositis, and follow-up implications: A Scottish population-based cohort study. *Br J Cancer* 2001; 85: 41-5. https://doi.org/10.1054/bjoc.2001.1699
25. OLDROYD A, ALLARD AB, CALLEN JP *et al*.: A systematic review and meta-analysis to inform cancer screening guidelines in idiopathic inflammatory myopathies. *Rheumatology* (Oxford) 2021; 60: 2615-28. https://doi.org/10.1093/rheumatology/keab166
26. ALLENBACH Y, KERAEN J, BOUVIER AM *et al*.: High risk of cancer in autoimmune necrotizing myopathies: Usefulness of myositis specific antibody. *Brain* 2016; 139: 2131-5. https://doi.org/10.1093/brain/aww054
27. TRALLERO-ARAGUAS E, RODRIGO-PENDAS JA, SELVA-O'CALLAGHAN A *et al*.: Usefulness of anti-p155 autoantibody for diagnosing cancer-associated dermatomyositis: a systematic review and meta-analysis. *Arthritis Rheum* 2012; 64: 523-32. https://doi.org/10.1002/art.33379
28. CASCIOLA-ROSEN L, MAMMEN AL: Myositis autoantibodies. *Curr Opin Rheumatol* 2012; 24: 602-8. https://doi.org/10.1097/BOR.0b013e328358bd85
29. ICHIMURA Y, MATSUSHITA T, HAMAGUCHI Y *et al*.: Anti-NXP2 autoantibodies in adult patients with idiopathic inflammatory myopathies: Possible association with malignancy. *Ann Rheum Dis* 2012; 71: 710-3. https://doi.org/10.1136/annrheumdis-2011-200697
30. FREDI M, CAVAZZANA I, CERIBELLI A *et al*.: An Italian multicenter study on Anti-NXP2 antibodies: Clinical and serological associations. *Clin Rev Allergy Immunol* 2022; 63: 240-50. https://doi.org/10.1007/s12016-021-08920-y
31. FIORENTINO DF, CHUNG LS, CHRISTOPHERSTINE L *et al*.: Most patients with cancer-associated dermatomyositis have antibodies to nuclear matrix protein NXP-2 or transcription intermediary factor 1gamma. *Arthritis Rheum* 2013; 65: 2954-62. https://doi.org/10.1002/art.38093
32. IKEDA S, ARITA M, MISAKI K *et al*.: Incidence and impact of interstitial lung disease and malignancy in patients with polymyositis, dermatomyositis, and clinically amyopathic dermatomyositis: A retrospective cohort study. *Springerplus* 2015; 4: 240. https://doi.org/10.1186/s40064-015-1013-8
33. LU X, YANG H, SHU X *et al*.: Factors predicting malignancy in patients with polymyositis and dermatomyostis: A systematic review and meta-analysis. *Plos One* 2014; 9: e94128. https://doi.org/10.1371/journal.pone.0094128
34. BALLESTER B, MILARA J, CORTIJO J: Idiopathic pulmonary fibrosis and lung cancer: Mechanisms and molecular targets. *Int J Mol Sci* 2019; 20: 593. https://doi.org/10.3390/ijms20030593
35. LI J, YANG M, LI P, SU Z, GAO P, ZHANG J: Idiopathic pulmonary fibrosis will increase the risk of lung cancer. *Chin Med J* (Engl) 2014; 127: 3142-9.
36. LEE SW, JUNG SY, PARK MC, PARK YB, LEE SK: Malignancies in Korean patients with inflammatory myopathy. *Yonsei Med J* 2006; 47: 519-23. https://doi.org/10.3349/ymj.2006.47.4.519