A retrospective cohort study in Chinese patients with adult polymyositis and dermatomyositis: risk of comorbidities and subclassification using machine learning

J. Zhu[1], L. Wu[1,2], Y. Zhou[3], R. Wang[1], S. Chen[1], J. Zhao[4], S. Yu[4], S. Zheng[1], F. Xiao[1,2], H. Ren[4], M. Yang[4], J. Li[2]

[1]Department of Rheumatic & TCM Medical Center, Nanfang Hospital, Southern Medical University, Guangzhou, China; [2]Department of Internal Medicine of Traditional Chinese Medicine, School of Traditional Chinese Medicine, Southern Medical University, Guangzhou, China; [3]Department of Obstetrics, Guangdong Women and Children Hospital, Guangzhou, China; [4]Department of Rheumatology, Nanfang Hospital, Southern Medical University, Guangzhou, China.

Abstract

Objective

To identify the risk factors in Chinese patients with adult polymyositis and dermatomyositis for their comorbidities and explore a subclassification system.

Methods

Clinical records of 397 patients with idiopathic inflammatory myopathies were retrospectively reviewed. Logistic regression was used to identify potential risk factors for interstitial lung disease (ILD), other rheumatic diseases, and malignancy after bivariate analysis. Hierarchical clustering and decisional tree were utilised to identify subgroups and explore a subclassification system.

Results

A total of 119 polymyositis and 191 dermatomyositis patients were included. Anti-PM/Scl, anti-Ro52, anti-aminoacyl-tRNA synthetase and anti-MDA5 (adjusted odds ratios (AOR)=4.779, 1.917, 5.092 and 7.714 respectively) antibodies were risks ($p<0.05$), whereas overlapping malignancy was protective (AOR=0.107; $p=0.002$) for ILD across polymyositis, dermatomyositis and the total group. In subgroup models, Raynaud's phenomenon, arthralgia and semi-quantitative anti-nuclear antibody (AOR=51.233, 4.261, 3.047 respectively) were risks for other overlapping rheumatic diseases ($p<0.05$). For overlapping malignancy, male and anti-TIF1γ antibodies (AOR=2.533, 16.949) were risks ($p<0.05$), whereas disease duration and combination of ILD (AOR=0.954, 0.106) were protective in the total group ($p<0.05$); while anti-NXP2 antibodies were identified as risk factors (AOR=73.152; $p=0.038$) in polymyositis. Hierarchical clustering suggested a subclassification with 6 subgroups: malignancy overlapping dermatomyositis, classical dermatomyositis, polymyositis with severe muscle involvement, dermatomyositis with ILD, polymyositis with ILD, and overlapping of myositis with other rheumatic diseases.

Conclusion

Accompanying ILD, other rheumatic diseases and malignancy are strongly associated with clinical manifestation and myositis-specific or myositis-associated autoantibodies among Chinese polymyositis and dermatomyositis patients. The subclassification system proposed a more precise phenotype defining toward stratified treatments.

Key words

polymyositis; dermatomyositis, comorbidity, classification

Junqing Zhu, PhD
Lisheng Wu, MD
Yi Zhou, MD
Ran Wang, PhD
Shixian Chen, PhD
Jinjun Zhao, PhD
Shenyi Yu, PhD
Songyuan Zheng, MD
Fei Xiao, MD
Hao Ren, PhD
Min Yang, PhD
Juan Li, PhD

The work should be attributed to: Department of Internal Medicine of Traditional Chinese Medicine, School of Traditional Chinese Medicine, Southern Medical University, and Department of Rheumatic & TCM Medical Center, Nanfang Hospital, Southern Medical University, Guangzhou, China.

Please address correspondence to: Juan Li, Department of Internal Medicine of Traditional Chinese Medicine, School of Traditional Chinese Medicine, Southern Medical University, 1838 North of Guangzhou Avenue, 510510 Guangzhou, China. E-mail: lijuan@smu.edu.cn

Received on November 17, 2020; accepted in revised form on January 25, 2021.

Funding: the study was supported by the Natural Science Foundation of China [no. 81803932] and the Natural Science Foundation of Guangdong Province [no. 2018030310025 and 2017A030313868]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: none declared.

Introduction

Idiopathic inflammatory myopathy (IIM), also known as myositis, refers to a group of heterogeneous disorders including polymyositis (PM), dermatomyositis (DM), inclusion body myositis and immune-mediated necrotising myopathy (1). PM and DM were the primary subsets of IIM in adults with incidence from 1.16 to 19.0 and prevalence from 2.4 to 33.8 per million people per year (2). These disorders characterised by chronic muscle inflammation are frequently accompanied with extramuscular manifestations, leading to an increased mortality rate (3). Different forms of myositis share clinical, pathological, and genetic characteristics, but differ in frequency of systemic involvement, degree of overlap with other rheumatic diseases or malignancy, and response to therapies and prognosis.

Multi-organ involvement and overlapping other diseases are characteristics also prognostic factors of myositis. Interstitial lung disease (ILD) is a common feature of lung involvement in PM and DM (4), with a prevalence rate of 5% to 65% in myositis, depending on disease subtype and risks in different populations (5, 6). Although anti-aminoacyl-tRNA synthetase antibodies (anti-ARS antibodies, commonly referred to as antisynthetase autoantibodies) have been considered a risk factor for ILD, severity among different anti-ARS antibodies varied (7). Other risk factors for myositis with ILD, such as specific clinical manifestations, other overlapping diseases, myositis-associated autoantibodies (MAAs) and other myositis-specific autoantibodies (MSAs) (8) have not been evaluated comprehensively. Meanwhile, studies with IIM examining the relationship between MSAs, MAAs and clinical manifestations have been more common in Caucasian populations (9, 10) with few in the Chinese patients (11), and distinct difference existed even in the previously considered "homogeneous" Japanese and Chinese populations (12). A study has demonstrated that MAAs may be risk factors for myositis with other overlapping rheumatic diseases (13), but there are few publications on factors contributing to clinical manifestation. Regarding overlapping with malignancy, studies demonstrated that individuals with positive anti-nuclear matrix protein 2 (anti-NXP2) autoantibodies had an increased risk of cancer (14, 15), while others found no such correlation (16, 17). And studies published were mainly on the relationship between anti-TIF1γ antibodies and cancer (18-20) in IIM. Therefore, it is necessary to systematically evaluate the relevant risk factors of myositis for ILD, other rheumatic diseases, and malignancy for better clinical vigilance.

Phenotype, pathogenesis, and prognosis vary due to multi-organ involvement and comorbidities. With the clinical application of MSAs, a new classification system for myositis was explored to reduce confusion between subgroups (21). But it is far from showing the full picture of myositis due to high heterogeneity. Further exploring the subclassification of myositis is critical. This study is to describe the clinical characteristics of adult PM and DM (the major subsets of IIM), assess risk factors for ILD and overlap of other diseases, and explore a new subclassification system by cluster analysis with multiple variables.

Materials and methods

Clinical data of 397 patients with IIM were collected from Nanfang Hospital in Guangdong, China, from October 2016 to July 2019. The 2017 European League Against Rheumatism/American College of Rheumatology classification criteria was applied (22). Patients fulfilled the classification criteria for PM or DM were included. Those younger than 16 years old, with any other type of myositis (*e.g.* inclusion body myositis or immune-mediated necrotising myopathy) or those missing important clinical data were excluded. All patients had provided consent forms, which were approved by the Institutional Medical Ethics Review Board of Nanfang Hospital (29160510).

The following clinical variables were collected: age, sex, disease duration, muscle involvement, skin involvement, arthralgia, Raynaud's phenomenon, lung involvement including ILD, overlapping other rheumatic diseases

Table I. Characteristics of patients with adult PM and DM included in the study.

	Total[&]	PM	DM	Test[#]	*p*-value
Number of cases (n)	310	119	191	NA	NA
Age (years)	49.170 ± 13.033	48.647 ± 13.383	49.497 ± 12.836	0.558	0.577
Male/Female (n)	111/199	35/84	76/115	3.436	0.064
Disease duration (months)	19.209 ± 44.802	24.596 ± 57.056	15.853 ± 34.831	2.705	0.007
Muscle involvement					
Muscle strength of upper limbs (0-V, n)	0/1/12/77/147/73	0/0/5/37/52/25	0/1/7/40/95/48	-1.632	0.103
Muscle strength of lower limbs (0-V, n)	0/2/21/83/138/66	0/1/12/37/44/25	0/1/9/46/94/41	-1.779	0.075
Muscle strength of flexor cervicalis (0-V, n)	1/0/8/34/97/170	1/0/5/21/45/47	0/0/3/13/52/123	-4.625	<0.001
Swallowing muscles (n, %)	91 (29.4%)	21 (17.6%)	70 (36.6%)	12.766	<0.001
Vocal muscles (n, %)	30 (9.7%)	8 (6.7%)	22 (11.5%)	1.929	0.165
Respiratory muscles (n, %)	31 (10.0%)	14 (11.8%)	17 (8.9%)	0.668	0.414
Myalgia (n, %)	177 (57.1%)	66 (55.5%)	111 (58.1%)	0.211	0.646
Skin involvement					
Heliotrope rash (n, %)	146 (47.1%)	NA	146 (76.4%)	NA	NA
Shawl sign (n, %)	67 (21.6%)	NA	67 (35.1%)	NA	NA
Anterior cervical V-shaped rash (n, %)	121 (39.0%)	NA	121 (63.4%)	NA	NA
Gottron's sign and papules (n, %)	87 (28.1%)	NA	87 (45.5%)	NA	NA
Mechanic's hands (n, %)	21 (6.8%)	NA	21 (11.0%)	NA	NA
Lesion surrounding the nails (n, %)	16 (5.2%)	NA	16 (8.4%)	NA	NA
Skin lesion ulcer (n, %)	46 (14.8%)	NA	46 (24.1%)	NA	NA
Holster's sign (n, %)	47 (15.2%)	NA	47 (24.6%)	NA	NA
Arthralgia (n, %)	89 (28.7%)	40 (33.6%)	49 (25.7%)	2.269	0.132
Raynaud's phenomenon (n, %)	25 (8.1%)	10 (8.4%)	15 (7.9%)	0.030	0.863
Lung involvement					
ILD (n, %)	128 (41.3%)	54 (45.4%)	74 (38.7%)	1.331	0.249
PAH (n, %)	26 (8.4%)	18 (15.1%)	8 (4.2%)	11.415	0.001
Lung infection (n, %)	151 (48.7%)	54 (45.4%)	97 (50.8%)	0.858	0.354
Respiratory failure (n, %)	34 (11.0%)	15 (12.6%)	19 (9.9%)	0.530	0.467
Overlapping other rheumatic diseases (n, %)	50 (16.1%)	27 (22.7%)	23 (12.0%)	6.144	0.013
RA/AS/SLE/PSS/SSC/MCTD (n)	13/1/16/10/14/2	9/0/7/10/4/1	4/1/9/0/10/1	NA	NA
Overlapping malignancy (n, %)	58 (18.7%)	7 (5.9%)	51 (26.7%)	20.895	0.001
Before/Co-occurrence/ Afer PM&DM (n)	18/15/25	3/1/3	15/14/22	NA	NA
Semi-quantitative ANA (0/1/2/3+, n) (P%)	117/141/46/6 (62.3%)	35/60/20/4 (70.6%)	82/81/26/2 (57.1%)	2.405	0.016
Quantification of ANA (U/mL)	65.676 ± 91.759	95.210 ± 106.839	47.275 ± 75.614	4.232	<0.001
MAAs					
Anti-U1-RNP antibodies (0/1/2/3+, n) (P%)	292/5/5/8 (5.8%)	112/4/1/2 (5.9%)	180/1/4/6 (5.8%)	-0.011	0.991
Anti-PM/Scl antibodies (0/1/2/3+, n) (P%)	295/4/8/3 (4.8%)	116/2/1/0 (2.5%)	179/2/7/3 (6.3%)	-1.535	0.125
Anti-Ro52 antibodies (0/1/2/3+, n) (P%)	180/32/18/80 (41.9%)	57/11/7/44 (52.1%)	123/21/11/36 (35.6%)	3.334	0.001
Anti-Ku antibodies (0/1/2/3+, n) (P%)	303/3/2/2 (2.3%)	115/1/1/2 (3.4%)	188/2/1/0 (1.6%)	1.048	0.295
Anti-ARS antibodies (n, P%)	82 (26.5%)	45 (37.8%)	37 (19.4%)	12.820	<0.001
Anti-Jo1 antibodies (0/1/2/3+, n) (P%)	260/13/4/33 (16.1%)	89/3/1/26 (25.2%)	171/10/3/7 (10.5%)	3.730	<0.001
Anti-PL7 antibodies (0/1/2/3+, n) (P%)	296/6/5/3 (4.5%)	112/3/2/2 (5.9%)	184/3/3/1 (3.7%)	0.920	0.358
Anti-EJ antibodies (0/1/2/3+, n) (P%)	303/2/2/3 (2.3%)	115/1/1/2 (3.4%)	188/1/1/1 (1.6%)	1.035	0.300
Anti-OJ antibodies (0/1/2/3+, n) (P%)	304/4/1/1 (1.9%)	117/2/0/0 (1.7%)	187/2/1/1 (2.1%)	-0.268	0.789
Anti-PL12 antibodies (0/1/2/3+, n) (P%)	305/3/2/0 (1.6%)	117/1/1/0 (1.7%)	188/2/1/0 (1.6%)	0.078	0.938
Other MSAs					
Anti-Mi-2 antibodies (0/1/2/3+, n) (P%)	294/11/4/1 (5.2%)	115/3/1/0 (3.4%)	179/8/3/1 (6.3%)	-1.137	0.256
Anti-TIF1γ antibodies (0/1/2/3+, n) (P%)	287/13/6/4 (7.4%)	114/3/1/1 (4.2%)	173/10/5/3 (9.4%)	-1.705	0.088
Anti-NXP2 antibodies (0/1/2/3+, n) (P%)	303/3/3/1 (2.3%)	116/1/2/0 (2.5%)	187/2/1/1 (2.1%)	0.246	0.806
Anti-MDA5 antibodies (0/1/2/3+, n) (P%)	286/10/9/5 (7.7%)	117/1/1/0 (1.7%)	169/9/8/5 (11.5%)	-3.157	0.002
Anti-SAE1 antibodies (0/1/2/3+, n) (P%)	307/1/1/1 (1.0%)	119/0/0/0 (0.0%)	188/1/1/1 (1.6%)	-1.372	0.170
Anti-SRP antibodies (0/1/2/3+, n) (P%)	297/5/3/5 (4.2%)	108/4/2/5/ (9.2%)	189/1/1/0 (1.0%)	3.510	<0.001
Relevant serum index					
White blood count (×10^9/L)	8.369 ± 4.879	9.188 ± 4.748	7.858 ± 4.902	3.379	0.001
Neutrophil count (×10^9/L)	6.097 ± 4.229	6.671 ± 4.616	5.739 ± 3.939	2.358	0.018
Percentage of neutrophils (%)	70.229 ± 11.781	69.267 ± 12.464	70.828 ± 11.326	1.135	0.257
Haemoglobin concentration (g/L)	122.023 ± 20.731	126.118 ± 17.992	119.471 ± 21.929	2.939	0.003
Platelet count (×10^9/L)	250.487 ± 99.502	270.546 ± 97.136	237.989 ± 99.160	3.083	0.002
Urinary protein (0/1/2/3+, n) (P%)	236/56/15/3(23.9%)	88/23/8/0(26.1%)	148/33/7/3(22.5%)	0.713	0.476
Quantification of urinary protein (g/24h)	0.369 ± 1.384	0.345 ± 0.711	0.384 ± 1.672	1.257	0.209
Urinary red blood cell count (/ul)	50.325 ± 393.980	20.178 ± 59.190	69.108 ± 499.336	0.868	0.386
Erythrocyte sedimentation rate (mm/1h)	33.565 ± 29.479	32.706 ± 30.096	34.099 ± 29.154	-0.808	0.419
C-reactive protein (mg/L)	19.187 ± 33.355	19.703 ± 39.944	18.865 ± 28.604	-1.548	0.122
Urea nitrogen (mmol/L)	4.896 ± 2.758	5.112 ± 3.451	4.760 ± 2.220	0.592	0.554
Creatinine (umol/L)	59.552 ± 53.644	64.479 ± 82.871	56.482 ± 19.755	-1.733	0.083

	Total[&]	PM	DM	Test[#]	*p*-value
Uric acid (umol/L)	331.261 ± 110.167	352.849 ± 114.878	317.812 ± 105.207	2.561	0.010
Cystatin C (mg/L)	1.229 ± 0.753	1.267 ± 1.117	1.204 ± 0.383	-1.063	0.288
Glutamic-oxalacetic transaminase (U/L)	158.732 ± 211.189	193.084 ± 275.030	137.330 ± 156.070	1.231	0.218
Glutamic-pyruvic transaminase (U/L)	102.571 ± 140.055	133.353 ± 187.963	83.393 ± 94.846	3.080	0.002
Lactate dehydrogenase (U/L)	663.384 ± 701.818	854.328 ± 934.933	544.419 ± 470.228	3.000	0.003
Alpha-hydroxybutyrate dehydrogenase (U/L)	437.010 ± 323.753	515.437 ± 346.648	388.147 ± 299.304	3.404	0.001
Creatine kinase (U/L)	2734.400 ± 4576.894	4158.765 ± 6080.811	1846.969 ± 3006.897	5.630	<0.001
Creatine kinase isoenzyme (U/L)	115.926 ± 188.206	177.353 ± 247.949	77.654 ± 124.858	5.116	<0.001
Cardiac troponin I (ng/mL)	0.152 ± 0.501	0.159 ± 0.464	0.147 ± 0.524	0.804	0.421
Myoglobin (ng/mL)	451.487 ± 405.135	616.218 ± 459.987	348.853 ± 328.544	5.410	<0.001
Brain natriuretic peptide (pg/mL)	812.529 ± 3568.552	1104.960 ± 4117.530	633.429 ± 3184.430	1.339	0.181
Serum total protein (g/L)	62.442 ± 8.819	65.038 ± 9.699	60.825 ± 7.823	4.923	<0.001
Albumin (g/L)	33.176 ± 6.095	34.846 ± 6.441	32.135 ± 5.640	3.831	<0.001
Globulin (g/L)	29.499 ± 6.061	30.699 ± 6.597	28.751 ± 5.590	2.362	0.018
Complement C3 (g/L)	1.085 ± 0.27	1.079 ± 0.288	1.089 ± 0.409	1.065	0.287
Complement C4 (g/L)	0.269 ± 0.116	0.250 ± 0.119	0.280 ± 0.113	-2.663	0.008
Total complement activity CH50 (U/mL)	51.858 ± 11.077	53.796 ± 10.802	50.651 ± 11.103	2.197	0.028
Drug application					
GCs (mg)	59.710 ± 34.377	72.479 ± 38.598	51.754 ± 28.827	4.252	<0.001
HGGs (n, %)	106 (34.2%)	46 (38.7%)	60 (31.4%)	1.709	0.191
DMARDs (n, %)	223 (71.9%)	94 (79.0%)	129 (67.5%)	4.763	0.029
NSAIDs (n, %)	109 (35.2%)	44 (37.0%)	65 (34.0%)	0.279	0.598
Anti-pulmonary fibrosis drugs (n, %)	59 (19.0%)	25 (21.0%)	34 (17.8%)	0.489	0.484
Drugs to improve microvascular circulation (n, %)	83 (26.8%)	43 (36.1%)	40 (20.9%)	8.631	0.003

Note: PM: polymyositis; DM: dermatomyositis; n: numbers; 0-V: muscle strength 0, I, II, III, IV, V; +, positive; ILD: interstitial lung disease; PAH: pulmonary artery hypertension; RA: rheumatoid arthritis; AS: ankylosing spondylitis; SLE: systemic lupus erythematosus; PSS: primary Sjogren's syndrome; SSC: systemic sclerosis; MCTD: mixed connective tissue disease; ANA: antinuclear antibody; MAAs: myositis-associated autoantibodies; ARS: aminoacyl-tRNA synthetase (commonly referred to as antisynthetase autoantibodies); MSAs: myositis-specific autoantibodies; GCs: glucocorticoids; HGGs: human gamma globulin; DMARDs: disease-modifying anti-rheumatic drugs; NSAIDs: non-steroidal anti-inflammatory drugs; Values are given as the numbers: percentage (%): or mean ± standard deviation; P%: percentage of the positive antibodies; &: patients with PM and DM; #: Compared between PM and DM: test statistic (t value or chi-square value) or standardised test statistic (z value) were reported; NA: not available.

and malignancy, semi-quantitative and quantification of anti-nuclear antibody (ANA), MAAs, anti-ARS antibodies, other MSAs, relevant serum index and drug administration (Table I). The Medical Research Council muscle strength grading system was used to measure muscle strength (grade 0, no contraction; grade 1, flicker or trace of contraction; grade 2, active movement with gravity eliminated; grade 3, active movement against gravity; grade 4, active movement against gravity and resistance; grade 5, normal power) (23). MAAs and MSAs were measured by line blot.

Statistical analyses were conducted using SPSS version 20 (SPSS Inc., Chicago, IL, USA). Continuous data were presented as means and standard deviations, and frequency data as numbers (n) and percentages (%). Kolmogorov-Smirnov and Levene's test were used to evaluate assumptions of normality and homogeneity of variance, respectively, of continuous data; a p-value <0.1 was considered statistically significant. These assumptions satisfied, Student's t test or one-way analysis was used to evaluate the differences. For non-normally distributed continuous data, Mann-Whitney U test or Kruskall-Wallis test was applied. For frequency data, Chi-square, Fisher's exact test with Monte Carlo simulation, or Mann-Whitney U-tests was used for the comparison of unordered and ordinal categorical variables, respectively. Test statistic (t, F, U, H, Z, or χ^2 value) and associated p-value were reported. Bivariate associations between clinical characteristics of patients and the risk factors for ILD, other rheumatic diseases and malignancy were analysed within the PM and DM groups separately. Odds ratios and 95% confidence intervals (CI) were reported. Logistic regression was used to identify the independent effect of each potential risk factor adjusting for others in the model. Variables with a p-value <0.05 in the bivariate analysis were included in the logistic regression models, for which adjusted odds ratios (AOR) and 95% CI were reported.

Multiple factor analysis (MFA), hierarchical clustering on principal components (HCPC), and classification and regression trees (CART) were performed with R version 3.6.2 (R Core Team, 2019) (24). Seventy-one discriminant variables were selected and organised in 18 groups (Supplementary File 1) according to their clinical significance to perform MFA. Dimensions kept for HCPC from MFA was determined at the point beyond which the remaining eigenvalues were all relatively small and of comparable size (25). HCPC was applied on the result of MFA to aggregate patients in clusters using package FactoMineR (26). The hierarchical tree was automatically cut at the suggested level. Lastly, CART was performed with package rpart to generate a decisional algorithm to predict clusters of patients based on variables selected in MFA (27). All patients were split randomly into training set (80%) for model building and test set (20%) for model evaluation. The optimal complexity parameter (cp) val-

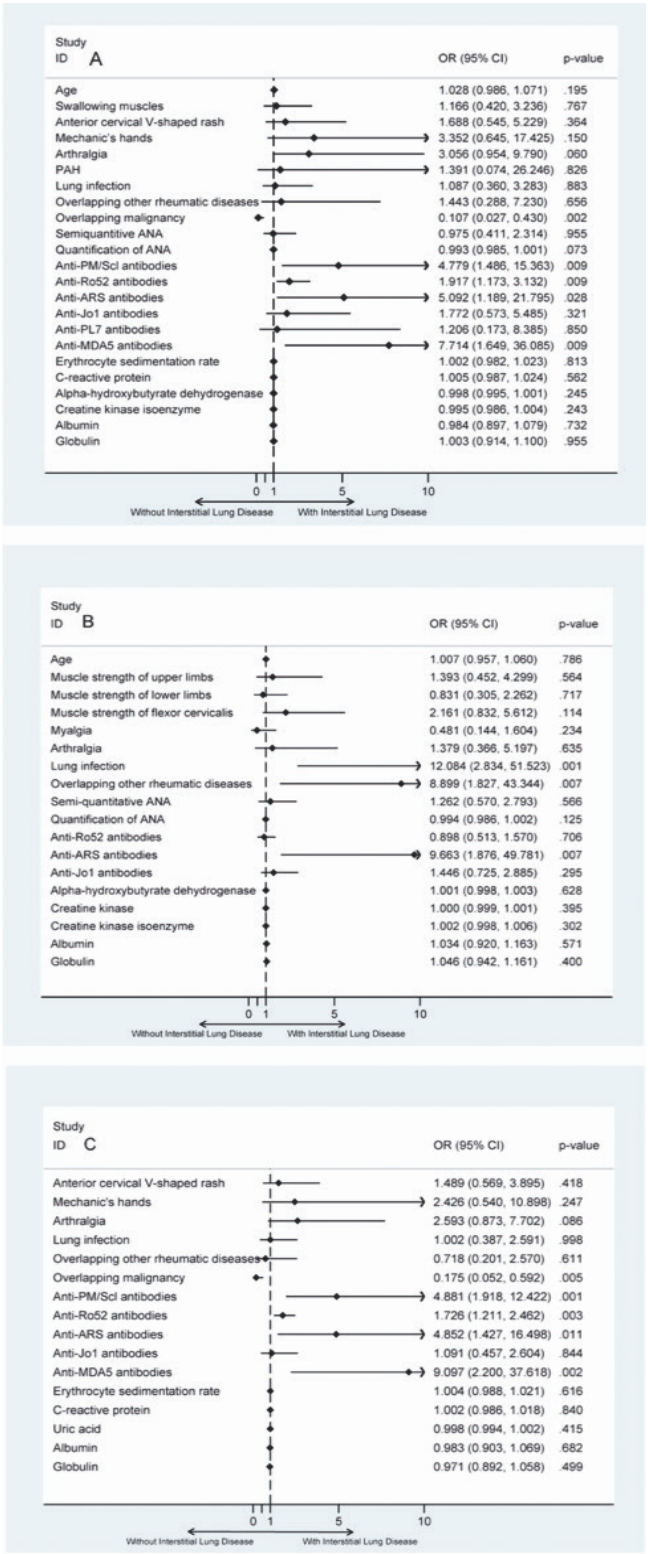

Fig. 1. The risk factors of interstitial lung disease in PM and DM
The logistic regression analysis was performed to identify the effect of each potential risk factor adjusted for others. The adjusted odds ratios (OR) and its 95% confidence interval (CI) for interstitial lung disease were reported. A, total patients with PM and DM; B, patients with PM; C, patients with DM. PM, polymyositis; DM, dermatomyositis; ILD, interstitial lung disease; PAH, pulmonary artery hypertension; ANA, antinuclear antibody; ARS, aminoacyl-tRNA synthetase (commonly referred to as antisynthetase autoantibodies).

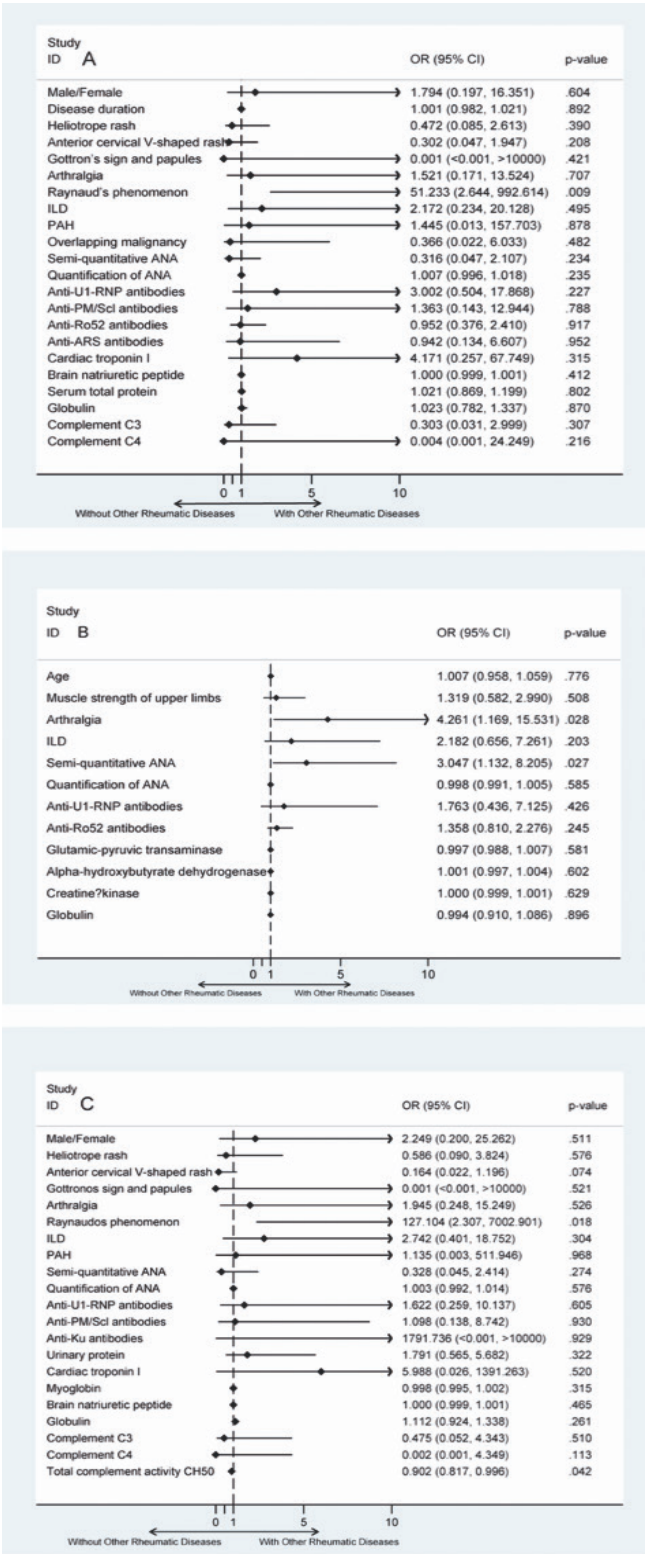

Fig. 2. The risk factors of overlapping other rheumatic diseases in PM and DM
The logistic regression analysis was performed to identify the effect of each potential risk factor adjusted for others. The adjusted odds ratios (OR) and its 95% CI for other rheumatic diseases were reported. A, total patients with PM and DM; B, patients with PM; C, patients with DM. PM, polymyositis; DM, dermatomyositis; ILD, interstitial lung disease; PAH, pulmonary artery hypertension; ANA, antinuclear antibody; ARS, aminoacyl-tRNA synthetase (commonly referred to as antisynthetase autoantibodies).

Table II. Types of overlapping malignancy in patients with adult PM and DM.

Overlapping of malignancy	PM (n=119)				DM (n=191)			
	Total (n)	Co-occurrence	Occurrence before IIM	Occurrence after IIM	Total (n)	Co-occurrence	Occurrence before IIM	Occurrence after IIM
Nasopharyngeal carcinoma	1	NA	1	NA	20	8	5	7
Lung cancer	0	NA	NA	NA	7	2	1	4
Thymic carcinoma	1	1	NA	NA	0	NA	NA	NA
Laryngeal cancer	0	NA	NA	NA	1	NA	1	NA
Oesophageal cancer	1	NA	NA	1	1	1	NA	NA
Gastric cancer	0	NA	NA	NA	2	1	1	NA
Colorectal and Rectal cancer	1	NA	NA	1	4	1	2	1
Cholangio carcinoma	0	NA	NA	NA	2	1	1	NA
Primary Hepatocellular carcinoma	0	NA	NA	NA	2	2	NA	NA
Breast cancer	1	NA	1	NA	3	1	2	NA
Ovarian cancer	1	NA	NA	1	4	1	2	1
Cervical / Endometrial cancer	1	NA	1	NA	3	1	1	1
Lymphoma	0	NA	NA	NA	1	1	NA	NA
Leukaemia	0	NA	NA	NA	1	1	NA	NA
Total (n)	7	1	3	3	51	14	15	22

PM: polymyositis; DM: dermatomyositis; n: numbers: all data was shown by numbers; NA: not available.

ue for pruning the decisional tree was selected based on the minimum x-error (cross-validation error). Performance of CART was evaluated.

Results

Clinical characteristics of adult PM and DM patients

Clinical data of 397 IIM patients were collected. After excluding patients with inconsistent diagnoses and missing data, a total of 310 patients (119 PM and 191 DM) were included (Table I). The average disease duration was 24.596±57.056 and 15.853±34.831 months among PM and DM patients, respectively (Z=2.705, p=0.007). There were no significant differences in age, sex, arthralgia, and Raynaud's phenomenon between the two groups. DM patients had more severe involvement of flexor cervicalis (Z= -4.625, p<0.001) and swallowing muscle (χ^2=12.766, p<0.001). ILD (41.3%) and lung infection (48.7%) were the primary manifestations of lung involvement. Pulmonary artery hypertension presented in 15.1% of PM and 4.2% of DM patients (χ^2=11.415, p=0.001), and the cumulative frequencies of myositis overlapping other rheumatic diseases and malignancy were 16.1% and 18.7%, respectively. No significant differences existed in the total frequencies and distributions of overlapping other rheumatic diseases between groups. DM had more overlapping malignancy than PM (26.7% *vs.* 5.9%, χ^2=20.895, p=0.001); however, no significant difference was found between groups regarding the prevalence of malignancy or occurrence sequence of malignancy and myositis. Besides some relevant serum indices and drug administration, semi-quantification and quantification of ANA, anti-Ro52, anti-ARS (mainly anti-Jo1), anti-MDA5 and anti-SRP antibodies were significantly different between PM and DM (p<0.05).

Association between clinical characteristics and ILD

Overall, 128 (41.3%) patients had a diagnosis of ILD, including 54 PM and 74 DM patients (Table I). According to bivariate analysis, clinical variables were widely associated with ILD among the overall myositis patients (total myositis group) and subgroups with PM or DM. The unadjusted ORs were reported in Supplementary File 2 among the three groups.

All clinical characteristics that were significantly different based on bivariate analyses (Suppl. File 2) within each group were included in logistic regression. Anti-PM/Scl (AOR=4.779, 95% CI 1.486–15.363), anti-Ro52 (AOR=1.917, 95% CI, 1.173–3.132) anti-ARS (AOR=5.092, 95% CI 1.189–21.795) and anti-MDA5 (AOR=7.714, 95% CI 1.649–36.085; all p<0.05) antibodies were identified as risk factors, whereas overlapping malignancy (AOR=0.107, 95% CI 0.027–0.430, p=0.002) was identified as a protective factor for ILD among the total myositis, PM and DM groups (Fig. 1A). Among PM patients, the AOR was 12.084 (95% CI 2.834–51.523, p=0.001) for lung infection, 8.899 (95% CI 1.827–43.344, p=0.007) for other overlapping rheumatic diseases and 9.663 (95% CI 1.876–49.781, p=0.007) for anti-ARS antibodies (Fig. 1B), which were risk factors for ILD. Among DM patients, the AOR was 4.881 (95% CI 1.918–12.422, p=0.001) for anti-PM/Scl, 1.726 (95% CI 1.211–2.462, p=0.003) for anti-Ro52, 4.852 (95% CI 1.427–16.498, p=0.011) for anti-ARS and 9.097 (95% CI 2.200–37.618, p=0.002) for anti-MDA5 antibodies, which were risk factors for ILD, and the AOR was 0.175 (95% CI 0.052–0.592, p=0.005) for overlapping malignancy, which was identified as a protective factor (Fig. 1C).

Frequency of and risk factors for other overlapping rheumatic diseases

Fifty (16.1%) patients overlapped other rheumatic diseases, including 13 rheumatoid arthritis (RA), 1 ankylosing spondylitis (AS), 16 systemic lupus erythematosus (SLE), 10 primary Sjögren's syndrome (PSS), 14 systemic sclerosis (SSC), and 2 mixed connective tissue disease (MCTD), of whom 27 were PM and 23 were DM patients (Table I). Bivariate analysis demonstrated that certain clinical characteristics associated with an unadjust-

ed risk of other overlapping rheumatic diseases in the total myositis, PM and DM group (Supplementary File 2).

In logistic regression, a positive association was observed between Raynaud's phenomenon and other overlapping rheumatic diseases (AOR=51.233, 95% CI 2.644–992.614, p=0.009) in the total myositis group (Fig. 2A). In DM, same risk factor was identified (AOR=127.104, 95% CI 2.307–7002.901, p=0.018); while total complementary activity of CH50 (AOR=0.902, 95% CI 0.817–0.996, p=0.042) was observed as a protective factor (Fig. 2C). In PM, arthralgia (AOR=4.261, 95% CI 1.169–15.531, p=0.028) and semi-quantitative ANA (AOR=3.047, 95% CI 1.132–8.205, p=0.027) were risk factors for other overlapping rheumatic diseases when the model was adjusted for other variables (Fig. 2B).

Frequency of and risk factors for overlapping malignancy

Overall, 58 (18.7%) patients overlapped with malignancy, including 7 PM and 51 DM patients. Among DM, the most frequent (20 cases) overlapping malignancy was nasopharyngeal carcinoma. Common overlapping malignancy across groups were lung, colorectal and rectal, ovarian, breast and cervical or endometrial cancers. Details of types and occurrence sequence of overlapping malignancy with myositis are shown in Table II.

Based on bivariate analysis (Suppl. File 2), logistic regression indicated that male (AOR=2.533, 95% CI 1.014–6.330, p<0.05) and anti-TIF1γ antibodies (AOR=16.949; 95% CI 3.350–85.739, p<0.05) were risks, whereas disease duration of myositis (AOR=0.954, 95% CI 0.916–0.994, p<0.05) and combining ILD (AOR=0.106, 95% CI 0.032–0.356, p<0.05) were protectors for overlapping malignancy among total myositis when the model was adjusted for other variables (Fig. 3A). Subgroup analyses demonstrated that anti-NXP2 antibodies were risk factors for overlapping malignancy in PM after adjusting other variables (AOR=73.152; 95% CI 1.274–4198.774, p 0.038; Fig. 3B). In DM, relationships between overlapping malignancy and anti-TIF1γ anti-

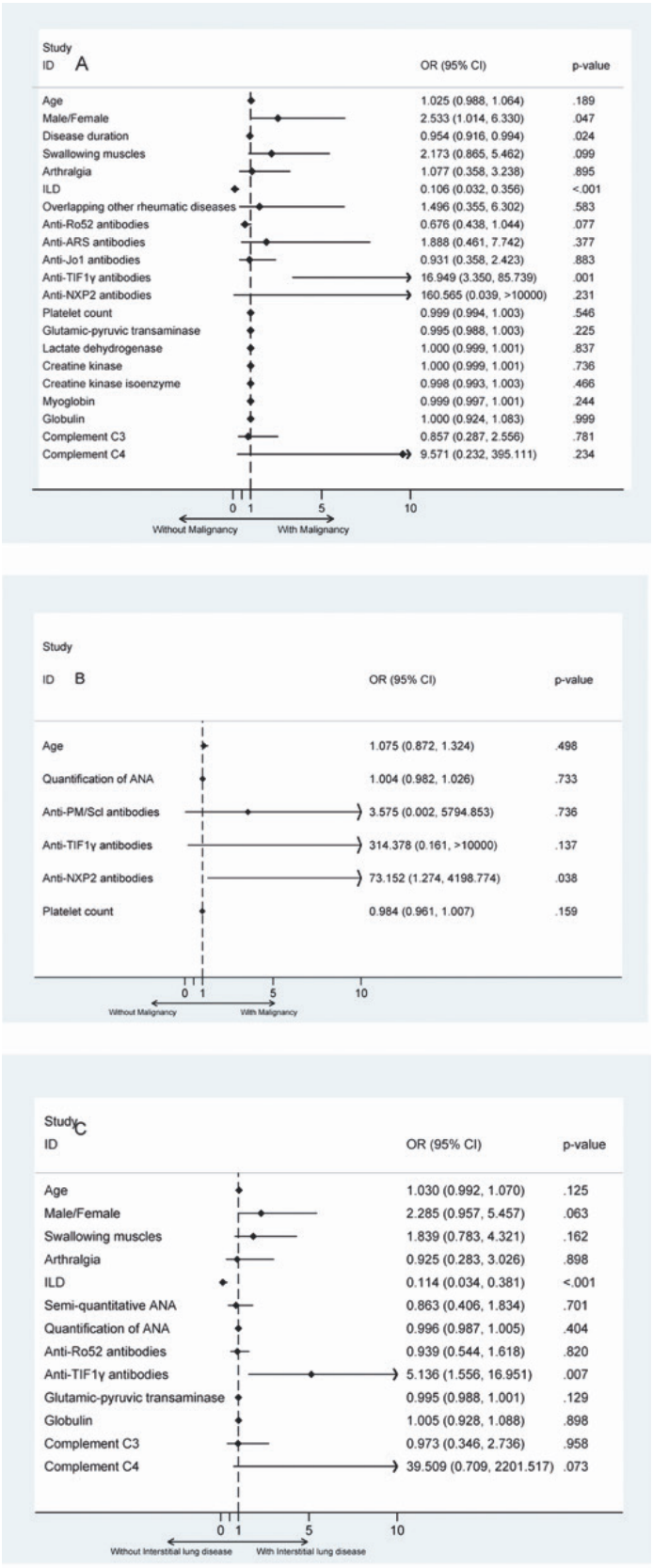

Fig. 3. The risk factors of overlapping malignancy in PM and DM.
The logistic regression analysis was performed to identify the effect of each potential risk factor adjusted for others. The adjusted odds ratios (OR) and its 95% CI for malignancy were reported. A, total patients with PM and DM; B, patients with PM; C, patients with DM. PM, polymyositis; DM, dermatomyositis; ILD, interstitial lung disease; ANA, antinuclear antibody.

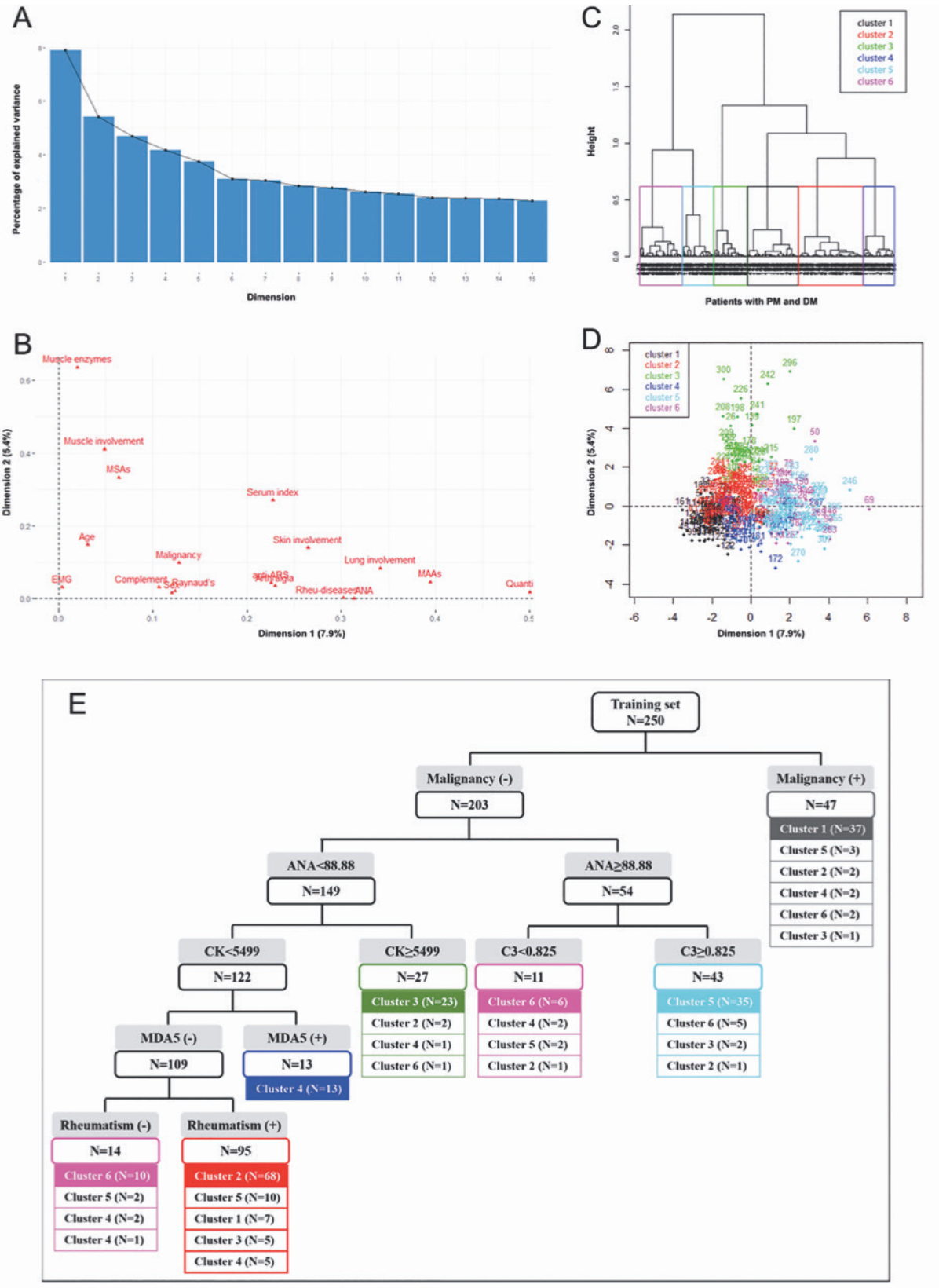

Fig. 4. Hierarchical clustering and final decisional tree for subclassification in adult PM and DM.
A, Scree plot of eigenvalues in descending order. The first 15 dimensions were shown. B, Correlation between grouped variables and the first two dimensions. Quantitative ANA and MAAs contributed the most to the first dimension. C, Dendrogram. X-axis indicated patients at the bottom. D, Factorial map. Patients plotted in the first two dimensions with each represented by a dot coloured correspondent to the patient's cluster. E, Decisional tree illustrated with patients in training set. PM, polymyositis; DM, dermatomyositis; ANA, quantification of ANA; CK, creatine kinase; C3, complement C3; MDA5, anti-MDA5 antibody; Rheumatism, overlapping other rheumatic diseases; +, positive; -, negative.

Table III. Characteristics of patients in clusters 1 to 6.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Test#	*p*-value
Number of cases (n)	55	93	38	30	65	29	NA	NA
Historical diagnosis (PM/DM, n)	1/54	33/60	26/12	3/27	44/21	12/17	NA	<0.001
Age (years)	55.95 ± 9.945	41.73 ± 10.983	44.66 ± 13.314	52.27 ± 12.649	55.72 ± 9.951	48.21 ± 15.776	17.152	<0.001
Male/Female (n)	37/18	31/62	9/29	15/15	15/50	4/25	NA	<0.001
Disease duration (months)	7.13 ± 7.991	29.66 ± 65.142	8 ± 9.792	5.7 ± 8.758	21.97 ± 42.626	31.17 ± 50.423	16.371	0.006
Muscle involvement								
Muscle strength of upper limbs (0-V, n)	0/0/4/11/23/17	0/0/0/12/50/31	0/1/6/27/3/1	0/0/0/6/22/2	0/0/2/14/33/16	0/0/0/7/16/6	NA	<0.001
Muscle strength of lower limbs (0-V, n)	0/0/5/10/25/15	0/0/0/19/44/30	0/1/15/16/6/0	0/0/0/9/21/0	0/1/1/21/24/18	0/0/0/8/18/3	NA	<0.001
Muscle strength of flexor cervicalis (0-V, n)	0/0/0/3/16/36	0/0/1/4/25/63	0/1/4/18/13/2	0/0/0/2/10/18	0/0/2/4/25/34	0/0/1/3/8/17	NA	<0.001
Swallowing muscles (n, %)	33 (60%)	18 (19.4%)	15 (39.5%)	8 (26.7%)	10 (15.4%)	7 (24.1%)	NA	<0.001
Vocal muscles (n, %)	6 (10.9%)	10 (10.8%)	6 (15.8%)	3 (10%)	3 (4.6%)	2 (6.9%)	NA	0.519
Respiratory muscles (n, %)	3 (5.5%)	5 (5.4%)	6 (15.8%)	11 (36.7%)	4 (6.2%)	2 (6.9%)	NA	<0.001
Myalgia (n, %)	26 (47.3%)	53 (57%)	29 (76.3%)	18 (60%)	33 (50.8%)	18 (62.1%)	NA	0.920
Skin involvement								
Heliotrope rash (n, %)	45 (81.8%)	51 (54.8%)	11 (28.9%)	20 (66.7%)	9 (13.8%)	10 (34.5%)	NA	<0.001
Shawl sign (n, %)	25 (45.5%)	17 (18.3%)	7 (18.4%)	8 (26.7%)	4 (6.2%)	6 (20.7%)	NA	<0.001
Anterior cervical V-shaped rash (n, %)	43 (78.2%)	38 (40.9%)	8 (21.1%)	18 (60%)	7 (10.8%)	7 (24.1%)	NA	<0.001
Gottron's sign and papules (n, %)	25 (45.5%)	34 (36.6%)	3 (7.9%)	19 (63.3%)	5 (7.7%)	1 (3.4%)	NA	<0.001
Mechanic's hands (n, %)	2 (3.6%)	5 (5.4%)	0 (0%)	8 (26.7%)	6 (9.2%)	0 (0%)	NA	0.010
Lesion surrounding the nails (n, %)	5 (9.1%)	4 (4.3%)	1 (2.6%)	5 (16.7%)	0 (0%)	1 (3.4%)	NA	0.014
Skin lesion ulcer (n, %)	13 (23.6%)	12 (12.9%)	1 (2.6%)	13 (43.3%)	4 (6.2%)	3 (10.3%)	NA	<0.001
Holster's sign (n, %)	14 (25.5%)	14 (15.1%)	3 (7.9%)	11 (36.7%)	2 (3.1%)	3 (10.3%)	NA	<0.001
Arthralgia (n, %)	4 (7.3%)	19 (20.4%)	4 (10.5%)	12 (40%)	37 (56.9%)	13 (44.8%)	NA	<0.001
Raynaud's phenomenon (n, %)	0 (0%)	4 (4.3%)	3 (7.9%)	0 (0%)	3 (4.6%)	15 (51.7%)	NA	<0.001
Lung involvement								
ILD (n, %)	8 (14.5%)	22 (23.7%)	8 (21.1%)	27 (90%)	46 (70.8%)	17 (58.6%)	NA	<0.001
PAH (n, %)	1 (1.8%)	2 (2.2%)	2 (5.3%)	1 (3.3%)	14 (21.5%)	6 (20.7%)	NA	<0.001
Lung infection (n, %)	24 (43.6%)	20 (21.5%)	14 (36.8%)	26 (86.7%)	53 (81.5%)	14 (48.3%)	NA	<0.001
Respiratory failure (n, %)	4 (7.3%)	3 (3.2%)	3 (7.9%)	14 (46.7%)	6 (9.2%)	4 (13.8%)	NA	<0.001
Overlapping other rheumatic diseases (n, %)	0 (0%)	2 (2.2%)	2 (5.3%)	4 (13.3%)	15 (23.1%)	27 (93.1%)	NA	<0.001
RA/AS/SLE/PSS/SSC/MCTD (n)	0/0/0/0/0/0	0/1/0/1/0/0	1/0/0/0/1/0	2/0/0/1/1/0	7/0/5/4/1/0	3/0/11/4/11/2	NA	0.014
Overlapping malignancy (n, %)	46 (83.6%)	2 (2.2%)	1 (2.6%)	2 (6.7%)	5 (7.7%)	2 (6.9%)	NA	<0.001
Before/Co-occurrence/After PM&DM (n)	13/20/13	1/1/0	1/0/0	0/0/2	3/1/1	0/0/2	NA	0.244
Semi-quantitative ANA (positive, n, %)	22 (40%)	46 (49.5%)	24 (63.2%)	12 (40%)	62 (95.4%)	27 (93.1%)	NA	<0.001
Quantification of ANA (U/mL)	20.888 ± 29.429	20.665 ± 29.807	33.995 ± 47.967	34.505 ± 61.078	169.364 ± 103.269	136.309 ± 117.68	124.17	<0.001
MAAs (positive, n, %)								
Anti-U1-RNP antibodies	4 (7.3%)	0 (0%)	1 (2.6%)	1 (3.3%)	1 (1.5%)	11 (37.9%)	NA	<0.001
Anti-PM/Scl antibodies	3 (5.5%)	1 (1.1%)	1 (2.6%)	0 (0%)	1 (1.5%)	9 (31%)	NA	<0.001
Anti-Ro52 antibodies	6 (10.9%)	20 (21.5%)	8 (21.1%)	20 (66.7%)	56 (86.2%)	20 (69%)	NA	<0.001
Anti-Ku antibodies	1 (1.8%)	1 (1.1%)	1 (2.6%)	0 (0%)	1 (1.5%)	3 (10.3%)	NA	0.132
Anti-ARS antibodies (positive, n, %)								
Anti-Jo1 antibodies	1 (1.8%)	6 (6.5%)	1 (2.6%)	0 (0%)	40 (61.5%)	2 (6.9%)	NA	<0.001
Anti-PL7 antibodies	0 (0%)	1 (1.1%)	0 (0%)	9 (30%)	4 (6.2%)	0 (0%)	NA	<0.001
Anti-EJ antibodies	0 (0%)	4 (4.3%)	1 (2.6%)	0 (0%)	0 (0%)	2 (6.9%)	NA	0.133
Anti-OJ antibodies	1 (1.8%)	2 (2.2%)	2 (5.3%)	0 (0%)	0 (0%)	1 (3.4%)	NA	0.395
Anti-PL12 antibodies	0 (0%)	1 (1.1%)	0 (0%)	1 (3.3%)	2 (3.1%)	1 (3.4%)	NA	0.439
Other MSAs (positive, n, %)								
Anti-Mi-2 antibodies	1 (1.8%)	11 (11.8%)	1 (2.6%)	1 (3.3%)	0 (0%)	2 (6.9%)	NA	0.011
Anti-TIF1γ antibodies	17 (30.9%)	0 (0%)	2 (5.3%)	0 (0%)	2 (3.1%)	2 (6.9%)	NA	<0.001
Anti-NXP2 antibodies	3 (5.5%)	0 (0%)	0 (0%)	0 (0%)	3 (4.6%)	1 (3.4%)	NA	0.085
Anti-MDA5 antibodies	1 (1.8%)	0 (0%)	1 (2.6%)	20 (66.7%)	0 (0%)	2 (6.9%)	NA	<0.001
Anti-SAE1 antibodies	0 (0%)	3 (3.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	NA	0.469
Anti-SRP antibodies	0 (0%)	0 (0%)	11 (28.9%)	0 (0%)	2 (3.1%)	0 (0%)	NA	<0.001
Relevant serum index								
White blood count (×10^9/L)	8.77 ± 6.272	6.782 ± 2.852	8.6 ± 3.993	9.118 ± 7.461	9.892 ± 4.523	8.206 ± 4.444	22.958	<0.001
Neutrophil count (×10^9/L)	6.169 ± 4.019	4.59 ± 2.445	6.26 ± 3.312	7.464 ± 7.177	7.64 ± 4.603	5.706 ± 3.875	26.893	<0.001
Percentage of neutrophils (%)	70.625 ± 10.2	66.058 ± 10.4	71.147 ± 10.381	76.787 ± 11.499	73.975 ± 13.166	66.466 ± 12.146	29.921	<0.001
Haemoglobin concentration (g/L)	117.47 ± 22.56	128.96 ± 18.196	125.58 ± 18.51	109.63 ± 27.579	119.86 ± 16.47	121.41 ± 20.488	25.074	<0.001
Platelet count (×10^9/L)	213.64 ± 90.582	237.88 ± 71.466	285.18 ± 94.631	235.67 ± 114.079	295.78 ± 123.306	229.14 ± 79.311	28.382	<0.001
Urinary protein (0/1/2/3+, n)	48/6/2/0	84/8/2/0	21/14/6/0	18/6/8/6	45/15/8/3	20/7/4/0	NA	<0.001
Quantification of urinary protein(g/24h)	0.157 ± 0.25	0.14 ± 0.226	1.028 ± 3.593	0.675 ± 1.103	0.389 ± 0.773	0.289 ± 0.428	27.774	<0.001
Urinary red blood cell count(/ul)	10.184 ± 12.549	49.155 ± 320.239	42.979 ± 111.477	233.233 ± 1123.599	22.166 ± 63.138	13.748 ± 16.814	9.455	0.092
Erythrocyte sedimentation rate(mm/1h)	25.95 ± 24.821	25.29 ± 23.8	22.11 ± 15.362	52.53 ± 29.072	51.4 ± 35.016	29.97 ± 29.897	55.473	<0.001
C-reactive protein (mg/L)	14.164 ± 20.078	8.109 ± 19.806	14.88 ± 27.971	37.873 ± 54.171	34.244 ± 43.481	16.82 ± 20.692	55.083	<0.001
Urea nitrogen (mmol/L)	5.254 ± 2.068	4.448 ± 1.678	4.999 ± 3.115	5.656 ± 5.307	5.094 ± 2.852	4.26 ± 1.974	9.403	0.094
Creatinine (umol/L)	62.53 ± 20.921	52.59 ± 18.146	63.82 ± 98.354	76.93 ± 118.008	58.83 ± 27.031	54.28 ± 19.962	18.803	0.002
Uric acid (umol/L)	340.98 ± 94.906	332.44 ± 104.588	353.95 ± 134.719	285 ± 115.259	325.43 ± 114.152	340.31 ± 98.811	8.405	0.135

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Test#	*p*-value
Cystatin C (mg/L)	1.288 ± 0.447	1.07 ± 0.391	1.03 ± 0.536	1.546 ± 1.792	1.374 ± 0.717	1.239 ± 0.434	25.031	<0.001
Glutamic-oxalacetic transaminase (U/L)	105.15 ± 114.56	103.52 ± 99.794	441.68 ± 392.175	130.9 ± 140.427	135.51 ± 158.25	147.59 ± 163.195	55.640	<0.001
Glutamic-pyruvic transaminase(U/L)	55.22 ± 47.648	73.97 ± 67.901	266.45 ± 280.44	95.33 ± 134.157	93.68 ± 89.158	96.83 ± 111.173	57.313	<0.001
Lactate dehydrogenase (U/L)	439.64 ± 385.3	443.42 ± 309.425	1527.53 ± 1057.676	609.07 ± 622.428	673.03 ± 707.526	695.41 ± 730.792	80.846	<0.001
Alpha-hydroxybutyrate dehydrogenase (U/L)	342.62 ± 251.129	341.72 ± 229.923	910.18 ± 366.843	349.5 ± 151.548	410.42 ± 258.558	451.86 ± 381.214	74.476	<0.001
Creatine kinase (U/L)	1177.8 ± 1755.331	1391.41 ± 1737.898	9635.21 ± 8515.879	1367.47 ± 1718.638	2819.29 ± 3675.437	2174.83 ± 2796.579	81.026	<0.001
Creatine kinase isoenzyme(U/L)	56.55 ± 78.441	60.73 ± 74.88	446.95 ± 332.395	79.83 ± 131.059	78.8 ± 84.683	92.34 ± 97.506	77.147	<0.001
Cardiac troponin I (ng/mL)	0.071 ± 0.17	0.041 ± 0.071	0.132 ± 0.306	0.298 ± 0.951	0.224 ± 0.693	0.371 ± 0.63	13.666	0.018
Myoglobin (ng/mL)	303.36 ± 281.903	338.71 ± 322.075	975.95 ± 481.707	267 ± 255.117	507.08 ± 375.156	473.14 ± 344.65	74.956	<0.001
Brain natriuretic peptide(pg/mL)	196.98 ± 326.932	174.86 ± 701.752	1250.21 ± 5557.422	1739.63 ± 5012.306	1045.61 ± 3117.653	1970 ± 6630.407	43.037	<0.001
Serum total protein (g/L)	58.524 ± 8.537	64.329 ± 7.434	62.258 ± 8.056	58.32 ± 6.834	63.922 ± 11.286	65.031 ± 6.211	32.932	<0.001
Albumin (g/L)	32.94 ± 5.978	35.806 ± 5.546	34.687 ± 5.598	28.53 ± 5.231	31.286 ± 5.905	32.252 ± 5.658	45.522	<0.001
Globulin (g/L)	25.584 ± 4.936	28.5 ± 5.194	27.571 ± 5.13	30.28 ± 4.356	33.552 ± 6.75	32.776 ± 4.768	67.559	<0.001
Complement C3 (g/L)	1.25 ± 0.529	1.09 ± 0.307	1.031 ± 0.265	1.035 ± 0.404	1.118 ± 0.301	0.84 ± 0.224	24.979	<0.001
Complement C4 (g/L)	0.321 ± 0.101	0.259 ± 0.099	0.272 ± 0.112	0.297 ± 0.114	0.279 ± 0.125	0.171 ± 0.076	40.860	<0.001
Total complement activity CH50 (U/mL)	48.973 ± 11.141	51.585 ± 10.832	51.416 ± 11.874	51.36 ± 11.673	57.195 ± 8.646	47.335 ± 11.248	24.692	<0.001

PM: polymyositis; DM: dermatomyositis; n: numbers; 0-V: muscle strength 0, I, II, III, IV, V; +, positive; ILD: interstitial lung disease; PAH: pulmonary artery hypertension; RA: rheumatoid arthritis; AS: ankylosing spondylitis; SLE: systemic lupus erythematosus; PSS: primary Sjögren's syndrome; SSC: systemic sclerosis; MCTD: mixed connective tissue disease; ANA: antinuclear antibody; MAAs: myositis-associated autoantibodies; ARS: aminoacyl-tRNA synthetase (commonly referred to as antisynthetase autoantibodies); MSAs: myositis-specific autoantibodies; Values are given as the numbers: percentage (%): or the mean ± standard deviation; #: Compared among group of Cluster 1-6; test statistic: F value or H value was reported; NA: Fisher's exact test with Monte Carlo simulation was used and test statistic was not available.

bodies (AOR=5.136, 95% CI 1.556–16.951; $p<0.05$) or combination of ILD (AOR=0.114, 95% CI 0.034–0.381; $p<0.05$) were demonstrated, which were consistent with findings among the total myositis group (Fig. 3C).

Exploration of subclassification for currently PM and DM based on machine learning

All 310 patients with 71 discriminant variables organised in 18 groups (Supplementary File 1) were included for analysis. Based on examination of the scree plot, the first five dimensions from the MFA results were applied for HCPC (Fig. 4A). Correlation between grouped variables and the first two dimensions was shown in Figure 4B. The hierarchical tree then suggested a partition of 6 clusters (Fig. 4C-D). Thirty-two categorical variables and 25 quantitative variables were discriminant in HCPC (Suppl. File 1).

Characteristics of patients and characterising variables of clusters 1 to 6 identified with HCPC were shown in Table III and Supplementary file 3. Cluster 1 (named malignancy overlapping DM) was characterised by composing of almost all DM patients, short disease duration, wide range of muscles involvement especially swallowing muscles, heliotrope rash, anterior cervical V-shaped rash, overlapping malignancy, and anti-TIF1γ antibodies positivity. Cluster 2 (named classical DM) was characterised with predominantly DM patients, long disease duration, wide muscles involvement, heliotrope rash, and anti-Mi-2 antibodies positivity. Cluster 3 (named PM with severe muscle involvement) was characterised by predominantly PM patients, short disease duration, wide muscles involvement especially respiratory muscles, anti-SRP antibodies positivity, and high relevant serum index (glutamic-oxalacetic transaminase, glutamic-pyruvic transaminase, lactate dehydrogenase, alpha-hydroxybutyrate dehydrogenase, creatine kinase, creatine kinase isoenzyme, and myoglobin). Cluster 4 (named DM with ILD) was characterised by mainly DM patients, short disease duration, wide muscles involvement especially respiratory muscles, heliotrope rash, anterior cervical V-shaped rash, Gottron's sign and papules, arthralgia, lung involvement (ILD, lung infection, and respiratory failure), anti-Ro52 antibodies positivity, anti-PL7 antibodies positivity, and anti-MDA5 antibodies positivity. Cluster 5 (named PM with ILD) was characterised by mainly PM patients, female, wide muscles involvement, arthralgia, lung involvement (ILD, PAH (pulmonary artery hypertension), and lung infection), high levels of ANA, anti-Ro52 antibodies positivity, and anti-Jo1 antibodies positivity. Cluster 6 (named overlapping of myositis with other rheumatic diseases) was characterised by female, long disease duration, wide muscles involvement, arthralgia, Raynaud's phenomenon, lung involvement (ILD and PAH), overlapping other rheumatic diseases, high levels of ANA, MAAs positivity (anti-U1-RNP antibodies, anti-PM/Scl antibodies, and anti-Ro52 antibodies), and low complement C3 and C4.

The final decisional tree was cut with split nodes 5 (cp value 0.023, x-error 0.394), with overlapping malignancy, quantification of ANA, creatine kinase, complement C3, Anti-MDA5 antibody and other overlapping rheumatic diseases in the final tree construction (Fig. 4E). Accuracy of the classification and regression trees model was 0.768 (95% CI 0.711–0.819) on training set and 0.633 (95%CI 0.499–0.754) on test set. Performance of the model by cluster was shown in Table IV. Balanced accuracy of each cluster in training and test set was about 80%, except in test set of clusters 5 and 6.

Discussion

IIM was initially considered mainly involving muscles and skin and were common with multi-organ involvement. In this study, we found that ILD and lung infection were the primary manifestations of lung involvement, which both affected more than 40% of the overall myositis patients. Incidence of ILD in these patients was similar to previous studies (5, 6). However, the

Table IV. Performance of classification and regression trees model by cluster.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Sensitivity	0.7872	0.8182	0.7158	0.6400	0.8519	0.50000	1.0000	1.00000	0.8140	0.4286	0.6400	0.37500
Specificity	0.9655	0.9592	0.9548	0.9429	0.9641	0.91071	0.9536	0.98182	0.9179	0.8113	0.9644	0.96154
PPV	0.8409	0.8182	0.9067	0.8889	0.7419	0.28571	0.5417	0.83333	0.6731	0.2308	0.6667	0.60000
NPV	0.9515	0.9592	0.8457	0.7857	0.9817	0.96226	1.0000	1.00000	0.9596	0.9149	0.9602	0.90909
Prevalence	0.1880	0.1833	0.3800	0.4167	0.1080	0.06667	0.0520	0.08333	0.1720	0.1167	0.1000	0.13333
Detection rate	0.1480	0.1500	0.2720	0.2667	0.0920	0.03333	0.0520	0.08333	0.1400	0.0500	0.0640	0.05000
Detection prevalence	0.1760	0.1833	0.3000	0.3000	0.1240	0.11667	0.0960	0.10000	0.2080	0.2167	0.0960	0.08333
Balanced accuracy	0.8764	0.8887	0.8353	0.7914	0.9080	0.70536	0.9768	0.99091	0.8659	0.6199	0.8022	0.66827

Training: training set; Test: test set; PPV: positive predictive value; NPV: negative predictive value.

proportion of lung infection was higher than that reported by Marie *et al*. (28) and Svensson *et al*. (29) among non-Chinese populations. The cumulative proportion of myositis overlapping other rheumatic diseases and malignancy was 16.1% and 18.7%, respectively. Malignancy occurred in 7/119 (5.9%) PM and 51/191 (26.7%) DM patients, consistent with previous understanding that DM is more likely to overlap malignancy (30). These complex manifestations of systemic involvement and comorbidities (*e.g*. other rheumatic diseases or malignancy) in IIM are possibly a result of shared immune mechanisms with different risk factors.

IIM associated ILD is a known contributor of excess mortality (31). There is consensus that certain MSAs are more likely to be associated with ILD, such as anti-ARS and anti-MDA5 antibodies (32). One study showed that different types of anti-ARS antibodies associated with ILD by the following order: PL-12 >PL-7 >Jo-1 (33). Our findings confirm that total anti-ARS antibodies are associated with risk of ILD in myositis and subgroups (PM and DM) in Chinese patients. Due to low prevalence of different anti-ARS antibodies, we did not find association between different types of anti-ARS antibodies and ILD. Anti-MDA5 antibodies have been linked with mucocutaneous ulcerations, ILD and mild muscle disease in adult or juvenile DM (34, 35). In patients among the myositis group and DM subgroup (not PM subgroup), anti-MDA5, anti-PM/Scl, and anti-Ro52 antibodies were associated with ILD. However, that MAAs associated with ILD may not be specific to myositis, but instead may be observed in the presence of overlapping diseases (36). This may explain why ILD but not MAAs among PM group was associated with other overlapping rheumatic diseases. In addition, this study revealed that overlapping malignancy was a protective factor for ILD among the total myositis and DM groups. One possible explanation is that prognosis of patients overlapping malignancy is poor (30). There may exists different immune mechanisms between myositis overlapping ILD and malignancy.

Other overlapping rheumatic diseases have been classified as overlapping myositis in some IIM criteria (3). Previous studies demonstrated that prevalence of other rheumatic diseases in myositis was low and primarily appeared as an overlapping syndrome (37, 38). In this study, 16.1% patients with myositis overlapped other rheumatic diseases (RA, AS, SLE, PSS, SSC, and MCTD), which were primarily related to collagen vascular diseases. Risks associated with other overlapping rheumatic diseases included Raynaud's, arthralgia, and semi-quantitative ANA in different groups of this study. As a result of low prevalence of specific overlapping rheumatic diseases, merged risk analysis may lead to bias.

Increased risk of cancer has been recognised for many years in myositis, particularly among DM (30, 39, 40). We identified 58 patients overlapping malignancy, including 7 (5.9%) with PM and 51 (26.7%) with DM. Myositis has not been associated with a certain malignancy, which may vary depending on the common cancers in a given population (15, 32). The most common malignancy in this study was nasopharyngeal carcinoma, which has a high incidence in southern China because of Epstein-Barr virus (41, 42). With regard to risk factors, anti-NXP2 antibodies (14), anti-TIF1γ (43) and anti-3-hydroxy-3-methylglutaryl coenzyme A reductase antibodies (44, 45) were associated with malignancy in IIM. However, a meta-analysis of 20 studies including 3,064 IIM patients, indicated that anti-NXP2 antibodies were not associated with malignancy (pooled OR=1.42, 95% CI 0.69–2.91) (46). In this study, anti-NXP2 and anti-TIF1γ antibodies were identified as risk factors for overlapping malignancy in PM and DM, respectively. These results are consistent with a previous study comprising a cohort of Chinese IIM patients (15). Moreover, we found that male was a risk factor for malignancy in PM and DM, whereas disease duration and combination of ILD were protective factors. Comprehensive screening for malignancy in patients with myositis is essential.

Among criteria proposed for IIM since 1970s, Bohan and Peter criteria in 1975 for PM and DM (47, 48) had been widely used for more than 40 years. Recently, the 2017 EULAR/ACR classification criteria for IIM and their subgroups have been proposed (22). The differences in the pathogenesis and in the clinical phenotype indicated the heterogeneity and complexity of IIM, which should be considered as a group of different diseases and not as a single disease (49). Lately, machine learning is being utilised to unravel the complexity of IIM (21, 50, 51). New classifica-

tion system has been explored, which included DM, inclusion body myositis, immune-mediated necrotising myopathy, and antisynthetase syndrome (21). With the progressing understanding of clinical characteristics other than muscles and skin involvement, and clinical application of MSAs, we explored a new subclassification for the currently defined PM and DM. Clustering approaches are effective unsupervised machine learning methods to identify homogeneous subgroups with similar attributes. Our hierarchical clustering on reduction of dimensions with MFA suggested 6 discriminate subgroups, which we named here as malignancy overlapping DM, classical DM, PM with severe muscle involvement, DM with ILD, PM with ILD, and overlapping of myositis with other rheumatic diseases according to the characteristics of grouped patients. The decisional tree highlighted the most relevant variables of each subgroup and predicted the group with minimal variables but meanwhile a decent performance. These findings, regarding DM, is similar to a previous study, in which patients with classical DM, older patients with malignancies, patients with ILD and patients with other connective tissue diseases were identified as distinct DM subgroups (51). As PM is now being considered a rare IIM subgroup with others especially connective tissue disease overlap myositis being the alternative diagnosis (52), it is sensible that the historically defined PM patients were regrouped into one of the 6 subgroups here. Our results, to some extent, indicated that myositis patients combined with ILD or other rheumatic diseases or malignancy not only had their own independent risk factors but also could be discriminated as independent subgroups for clinical reference.

This study has several limitations. First, the 2017 EULAR/ACR classification criteria were adopted; therefore, not all patients were pathologically diagnosed, which could have produced limitations on subclassification. Second, because of low prevalence of certain overlapping rheumatic diseases and malignancy in myositis, subgroup analysis was not performed. Third, high heterogeneity of myositis and limited sample size here may lead to misclassification in some clusters. Fourth, due to the limitation of the results at medical records, some clinical indicators were evaluated subjectively from patients (swallowing and respiratory muscles involvement) or not evaluated by gold standard and grading method (PAH). More objective and grading evaluation methods such as eating assessment tool (EAT)-10 or fibreoptic endoscopy for the evaluation of dysphagia are really needed (53). With discovery of new MSAs, further prospective studies comprising large samples and multiple centres would be valuable to accurately assess the risk of myositis systemic damage and comorbidities and be necessary to a more precise subclassification system.

Conclusions

Combining with ILD and overlapping other rheumatic diseases or malignancy are closely associated with clinical manifestations and MSAs or MAAs. To unravel the high heterogeneity of myositis, six clusters have been identified as an exploration of subclassification, including malignancy overlapping DM, classical DM, PM with severe muscle involvement, DM with ILD, PM with ILD, and overlapping of myositis with other rheumatic diseases. These findings could facilitate clinical management and contribute to refining classification of IIM.

Acknowledgements

The study was supported by the Natural Science Foundation of China (no. 81803932) and the Natural Science Foundation of Guangdong Province (no. 2018030310025 and 2017A030313868). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. LUNDBERG IE, DE VISSER M, WERTH VP: Classification of myositis. *Nat Rev Rheumatol* 2018; 14: 269-78.
2. MEYER A, MEYER N, SCHAEFFER M, GOTTENBERG JE, GENY B, SIBILIA J: Incidence and prevalence of inflammatory myopathies: A systematic review. *Rheumatology* (Oxford) 2015; 54: 50-63.
3. SELVA-O'CALLAGHAN A, PINAL-FERNANDEZ I, TRALLERO-ARAGUAS E, MILISENDA JC, GRAU-JUNYENT JM, MAMMEN AL: Classification and management of adult inflammatory myopathies. *Lancet Neurol* 2018; 17: 816-28.
4. HA YJ, LEE YJ, KANG EH: Lung involvements in rheumatic diseases: Update on the epidemiology, pathogenesis, clinical features, and treatment. *Biomed Res Int* 2018; 2018: 6930297.
5. MARIE I, HACHULLA E, CHERIN P *et al*.: Interstitial lung disease in polymyositis and dermatomyositis. *Arthritis Rheum* 2002; 47: 614-22.
6. FATHI M, DASTMALCHI M, RASMUSSEN E, LUNDBERG IE, TORNLING G: Interstitial lung disease, a common manifestation of newly diagnosed polymyositis and dermatomyositis. *Ann Rheum Dis* 2004; 63: 297-301.
7. MECOLI CA, CHRISTOPHER-STINE L: Management of interstitial lung disease in patients with myositis specific autoantibodies. *Curr Rheumatol Rep* 2018; 20: 27.
8. STUHLMULLER B, SCHNEIDER U, GONZALEZ-GONZALEZ JB, FEIST E: Disease specific autoantibodies in idiopathic inflammatory myopathies. *Front Neurol* 2019; 10: 438.
9. LILLEKER JB, VENCOVSKY J, WANG G *et al*.: The euromyositis registry: An international collaborative tool to facilitate myositis research. *Ann Rheum Dis* 2018; 77: 30-9.
10. BETTERIDGE Z, TANSLEY S, SHADDICK G *et al*.: Frequency, mutual exclusivity and clinical associations of myositis autoantibodies in a combined European cohort of idiopathic inflammatory myopathy patients. *J Autoimmun* 2019; 101: 48-55.
11. LI S, GE Y, YANG H *et al*.: The spectrum and clinical significance of myositis-specific autoantibodies in Chinese patients with idiopathic inflammatory myopathies. *Clin Rheumatol* 2019; 38: 2171-9.
12. CHEN Z, HU W, WANG Y, GUO Z, SUN L, KUWANA M: Distinct profiles of myositis-specific autoantibodies in Chinese and Japanese patients with polymyositis/dermatomyositis. *Clin Rheumatol* 2015; 34: 1627-31.
13. FREDI M, CAVAZZANA I, FRANCESCHINI F: The clinico-serological spectrum of overlap myositis. *Curr Opin Rheumatol* 2018; 30: 637-43.
14. ALBAYDA J, PINAL-FERNANDEZ I, HUANG W *et al*.: Antinuclear matrix protein 2 autoantibodies and edema, muscle disease, and malignancy risk in dermatomyositis patients. *Arthritis Care Res* (Hoboken) 2017; 69: 1771-6.
15. YANG H, PENG Q, YIN L *et al*.: Identification of multiple cancer-associated myositis-specific autoantibodies in idiopathic inflammatory myopathies: A large longitudinal cohort study. *Arthritis Res Ther* 2017; 19: 259.
16. WANG L, HUANG L, YANG Y *et al*.: Calcinosis and malignancy are rare in chinese adult patients with myositis and nuclear matrix protein 2 antibodies identified by an unlabeled immunoprecipitation assay. *Clin Rheumatol* 2018; 37: 2731-9.
17. BEST M, JACHIET M, MOLINARI N *et al*.: Distinctive cutaneous and systemic features associated with specific antimyositis antibodies in adults with dermatomyositis: A prospec-

tive multicentric study of 117 patients. *J Eur Acad Dermatol Venereol* 2018; 32: 1164-72.
18. OLDROYD A, SERGEANT JC, NEW P *et al*.: The temporal relationship between cancer and adult onset anti-transcriptional intermediary factor 1 antibody–positive dermatomyositis. *Rheumatology* 2018; 58: 650-5.
19. HIDA A, YAMASHITA T, HOSONO Y *et al*.: Anti-tif1-γ antibody and cancer-associated myositis: A clinicohistopathologic study. *Neurology* 2016; 87: 299-308.
20. DANI L, HOLMQVIST M, MARTÍNEZ MA *et al*.: Anti-transcriptional intermediary factor 1 gamma antibodies in cancer-associated myositis: A longitudinal study. *Clin Exp Rheumatol* 2020; 38: 67-73.
21. 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.
22. LUNDBERG IE, TJARNLUND A, BOTTAI M *et al*.: 2017 European League Against Rheumatism/American College of Rheumatology classification criteria for adult and juvenile idiopathic inflammatory myopathies and their major subgroups. *Arthritis Rheumatol* 2017; 69: 2271-82.
23. SEEDER L: Muscle strength grading. *Ann Emerg Med* 1983; 12: 407.
24. R core team. R: A language and environment for statistical computing. Vienna, Austria.; 2019.
25. CATTELL RB: The scree test for the number of factors. *Multivariate Behav Res* 1966; 1: 245-76.
26. LE S, JOSSE J, HUSSON F: Factominer: An r package for multivariate analysis. *J Stat Softw* 2008; 25: 1-18.
27. THERNEAU T, ATKINSON B. Rpart: Recursive partitioning and regression trees. 2019.
28. MARIE I, MENARD JF, HACHULLA E *et al*.: Infectious complications in polymyositis and dermatomyositis: A series of 279 patients. *Semin Arthritis Rheum* 2011; 41: 48-60.
29. SVENSSON J, HOLMQVIST M, LUNDBERG IE, ARKEMA EV: Infections and respiratory tract disease as risk factors for idiopathic inflammatory myopathies: A population-based case-control study. *Ann Rheum Dis* 2017; 76: 1803-8.
30. QIANG JK, KIM WB, BAIBERGENOVA A, ALHUSAYEN R: Risk of malignancy in dermatomyositis and polymyositis. *J Cutan Med Surg* 2017; 21: 131-6.
31. AMARAL SM, COGOLLO E, ISENBERG DA: Why do patients with myositis die? A retrospective analysis of a single-centre cohort. *Clin Exp Rheumatol* 2016; 34: 820-6.
32. MCHUGH NJ, TANSLEY SL: Autoantibodies in myositis. *Nat Rev Rheumatol* 2018; 14: 290-302.
33. PINAL-FERNANDEZ I, CASAL-DOMINGUEZ M, HUAPAYA JA *et al*.: A longitudinal cohort study of the anti-synthetase syndrome: increased severity of interstitial lung disease in black patients and patients with anti-pl7 and anti-pl12 autoantibodies. *Rheumatology* (Oxford) 2017; 56: 999-1007.
34. TANSLEY SL, BETTERIDGE ZE, GUNAWARDENA H *et al*.: Anti-mda5 autoantibodies in juvenile dermatomyositis identify a distinct clinical phenotype: a prospective cohort study. *Arthritis Res Ther* 2014; 16: R138.
35. FIORENTINO D, CHUNG L, ZWERNER J, ROSEN A, CASCIOLA-ROSEN L: The mucocutaneous and systemic phenotype of dermatomyositis patients with antibodies to mda5 (cadm-140): A retrospective study. *J Am Acad Dermatol* 2011; 65: 25-34.
36. LEGA JC, FABIEN N, REYNAUD Q *et al*.: The clinical phenotype associated with myositis-specific and associated autoantibodies: a meta-analysis revisiting the so-called antisynthetase syndrome. *Autoimmun Rev* 2014; 13: 883-91.
37. COLAFRANCESCO S, PRIORI R, GATTAMELATA A *et al*.: Myositis in primary Sjögren's syndrome: Data from a multicentre cohort. *Clin Exp Rheumatol* 2015; 33: 457-64.
38. IOANNOU Y, SULTAN S, ISENBERG DA: Myositis overlap syndromes. *Curr Opin Rheumatol* 1999; 11: 468-74.
39. HILL CL, ZHANG Y, SIGURGEIRSSON B *et al*.: Frequency of specific cancer types in dermatomyositis and polymyositis: A population-based study. *Lancet* 2001; 357: 96-100.
40. 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.
41. ZHAO FP, LIU X, CHEN XM *et al*.: Levels of plasma epstein-barr virus DNA prior and subsequent to treatment predicts the prognosis of nasopharyngeal carcinoma. *Oncol Lett* 2015; 10: 2888-94.
42. CHEN YP, CHAN ATC, LE QT, BLANCHARD P, SUN Y, MA J: Nasopharyngeal carcinoma. *Lancet* 2019; 394: 64-80.
43. OGAWA-MOMOHARA M, MURO Y, MITSUMA T *et al*.: Strong correlation between cancer progression and anti-transcription intermediary factor 1gamma antibodies in dermatomyositis patients. *Clin Exp Rheumatol* 2018; 36. 990-5.
44. KADOYA M, HIDA A, HASHIMOTO MM *et al*.: Cancer association as a risk factor for anti-hmgcr antibody-positive myopathy. *Neurol Neuroimmunol Neuroinflamm* 2016; 3: e290.
45. TINIAKOU E, PINAL-FERNANDEZ I, LLOYD TE *et al*.: More severe disease and slower recovery in younger patients with anti-3-hydroxy-3-methylglutaryl-coenzyme a reductase-associated autoimmune myopathy. *Rheumatology* (Oxford) 2017; 56: 787-94.
46. ZHONG L, YU Z, SONG H: Association of anti-nuclear matrix protein 2 antibody with complications in patients with idiopathic inflammatory myopathies: A meta-analysis of 20 cohorts. *Clin Immunol* 2019; 198: 11-8.
47. BOHAN A, PETER JB: Polymyositis and dermatomyositis (first of two parts). *N Engl J Med* 1975; 292: 344-7.
48. BOHAN A, PETER JB: Polymyositis and dermatomyositis (second of two parts). *N Engl J Med* 1975; 292: 403-7.
49. ZANFRAMUNDO G, TRIPOLI A, COMETI L *et al*.: One year in review 2020: Idiopathic inflammatory myopathies. *Clin Exp Rheumatol* 2021; 39: 1-12.
50. ENG SWM, OLAZAGASTI JM, GOLDENBERG A *et al*.: A clinically and biologically based subclassification of the idiopathic inflammatory myopathies using machine learning. *ACR Open Rheumatol* 2020; 2: 158-66.
51. ZHU H, WU C, JIANG N *et al*.: Identification of 6 dermatomyositis subgroups using principal component analysis-based cluster analysis. *Int J Rheum Dis* 2019; 22: 1383-92.
52. LOARCE-MARTOS J, LILLEKER JB, PARKER M, MCHUGH N, CHINOY H: Polymyositis: is there anything left? A retrospective diagnostic review from a tertiary myositis centre. *Rheumatology* (Oxford) 2021; 60: 3398-403.
53. GIANNINI M, FIORELLA ML, TAMPOIA M *et al*.: Long-term efficacy of adding intravenous immunoglobulins as treatment of refractory dysphagia related to myositis: A retrospective analysis. *Rheumatology* (Oxford) 2021: 60: 1234-42.