BioData Mining

RESEARCH

Open Access

Identifying heterogeneous subgroups of systemic autoimmune diseases by applying a joint dimension reduction and clustering approach to immunomarkers

Chia-Wei Chang[1†], Hsin-Yao Wang[2,7†], Wan-Ying Lin[2,3,4,8], Yu-Chiang Wang[4,9,10], Wei-Lin Lo[5], Ting-Wei Lin[2], Jia-Ruei Yu[2] and Yi-Ju Tseng[1,6*]

[†]Chia-Wei Chang and Hsin-Yao Wang contributed equally to this work.

*Correspondence:
Yi-Ju Tseng
yjtseng@nycu.edu.tw

Full list of author information is available at the end of the article

Abstract

Background The high complexity of systemic autoimmune diseases (SADs) has hindered precise management. This study aims to investigate heterogeneity in SADs.

Methods We applied a joint cluster analysis, which jointed multiple correspondence analysis and k-means, to immunomarkers and measured the heterogeneity of clusters by examining differences in immunomarkers and clinical manifestations. The electronic health records of patients who received an antinuclear antibody test and were diagnosed with SADs, namely systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and Sjögren's syndrome (SS), were retrieved between 2001 and 2016 from hospitals in Taiwan.

Results With distinctive patterns of immunomarkers, a total of 11,923 patients with the three SADs were grouped into six clusters. None of the clusters was composed only of a single SAD, and these clusters demonstrated considerable differences in clinical manifestation. Both patients with SLE and SS had a more dispersed distribution in the six clusters. Among patients with SLE, the occurrence of renal compromise was higher in Clusters 3 and 6 (52% and 51%) than in the other clusters ($p < 0.001$). Cluster 3 also had a high proportion of patients with discoid lupus (60%) than did Cluster 6 (39%; $p < 0.001$). Patients with SS in Cluster 3 were the most distinctive because of the high occurrence of immunity disorders (63%) and other and unspecified benign neoplasm (58%) with statistical significance compared with the other clusters (all $p < 0.05$).

Conclusions The immunomarker-driven clustering method could recognise more clinically relevant subgroups of the SADs and would provide a more precise diagnosis basis.

Keywords Autoimmune diseases, Immune markers, Cluster analysis, Disease heterogeneity

Introduction

Systemic autoimmune disease (SAD) is an umbrella term for autoimmune diseases that could affect all body systems and organs. SADs are typically divided into major categories, such as systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and Sjögren's syndrome (SS). However, patients with specific SADs present with considerably varying manifestations. Regarding the initial presentation of SLE, a molar rash, pleural effusion, or even septic shock can be the first manifestation of SLE [1]. Moreover, severity and prognosis considerably vary among patients with the same SAD [2]. Some patients meet sufficient criteria and receive a clear diagnosis of SAD, whereas other patients fail to meet criteria required for the diagnosis [3]. The high complexity of SADs has hindered their precise management.

To provide more meticulous treatments for SADs, detailed identification of heterogeneous SAD subgroups would be the key first step for developing tailored healthcare strategies. Regrouping or reclassification of various well-known diseases has attracted considerable interest. Diabetes, among the most prevalent diseases, is classified into four categories, not only type 1 and type 2, based on the traditional classification [4]. Myocardial infarction, another major disease with a high prevalence, is classified into five types based on pathological, clinical, and prognostic differences, and treatment strategies vary for these five types [5]. An update of the classification scheme enhances clinical usefulness and establishes an accurate diagnosis [6]. Due to their complex nature, the current classification scheme for SADs does not meet the clinical need.

Redefining SADs is an ongoing task in the rheumatology community. European and American committees established new classification criteria for SLE in 2019 [7]. In addition to experts' consensus, antinuclear antibody (ANA) testing and data-driven methods have been employed to refine the definition of SLE [7]. In fact, in the realm of disease diagnosis (e.g., autoimmune diseases, cancers), relying solely on individual biomarkers often proves inadequate [8, 9]. To counter this limitation, utilizing a biomarker panel consisting of multiple markers has shown substantial potential for enhancing diagnostic accuracy [10]. However, due to the intricate and implicit numerical patterns within such biomarker panels, the incorporation of machine learning and artificial intelligence technologies becomes pivotal in deciphering these specific disease patterns, thereby further amplifying diagnostic precision [11, 12]. On top of the approach, using numerous objective measurements (e.g., biomarker testing) with data-driven analytical methods has been advocated as a more adequate approach for redefining SADs. Using data-driven analytical methods, such as unsupervised clustering, studies have demonstrated that the heterogeneity of autoimmune diseases can be efficiently deconstructed to obtain clinically meaningful insights [13–19]. Hierarchical clustering with endoscopy and histology data was applied to identify subgroups of inflammatory bowel disease with differing colonic involvement [13] and subgroups of patients with anti-Ku syndrome developing different types of severe comorbidities using clinical features [18]. Consensus clustering and similarity network fusion algorithms were applied to identify subtypes in RA with RNA sequencing data [14] and subgroups of molecular disease mechanisms with OMIC dataset [15], respectively. Similar to our study's target, a previous study classified patients with particular SADs by performing cluster analysis based on principal components (PCs) derived from a multiple correspondence analysis (MCA), with laboratory measurements [19]. While the previous study observed dissimilarities in the positive

rates of autoantibodies and the frequencies of SADs between clusters, highlighting the importance of heterogeneity of SADs, the association between clusters and clinical characteristics remains unclear [19].

However, treating dimension reduction and clustering as separate processes may not define the informative subspace for genuine cluster structures. To ensure the most informative subspace for clustering is directly used, we applied a method that jointly performs dimension reduction and clustering. This method allows for more precise identification of clinically relevant clusters and their associations with clinical characteristics, addressing the abovementioned limitations. This study investigated heterogeneity in SADs with a method that jointed dimension reduction and clustering on immunomarkers that are frequently tested in SADs and are typically ordered during routine clinical visits, and evaluated heterogeneity based on clinical manifestations.

Method

Study population and setting

This study retrieved electronic medical records from the Chang Gung Research Database between 2001 and 2019, containing clinical data from three medical centres and five regional hospitals in Taiwan [20]. We included patients receiving a principal diagnosis of SADs (Supplement Document 1) between 2001 and 2016 and had at least one ANA test.

Patient inclusion algorithm

Figure 1 presents a flowchart of patient inclusion rules. We identified patients receiving at least two diagnoses of a specific SAD 30 days apart within 365 days [21]. Then, we included patients having at least one immunomarker result within 90 days before and 30 days after the first diagnosis date to minimise the bias from treatments that immunomarkers might be subject to. To have an adequate number of patients for each target SAD, we included only patients diagnosed with common SADs, namely SLE, SS, and RA, for the cluster analysis and excluded those with multiple SADs. Finally, patients having at least three years of follow-up from the first diagnosis were retained for a clinical implication analysis.

Data preprocessing and imputation

We used 33 immunomarkers as the data for cluster analysis. These immunomarkers are frequently tested, crucial for diagnosing and monitoring SADs, and typically ordered during routine clinical visits. This choice ensures the practical applicability of our model in real-world clinical environments. All the tests were conducted at Chang Gung Memorial Hospital, which has been accredited by The College of American Pathologists (CAP) since 2002.

We then converted the results of 33 immunomarkers (Supplement Document 2) into categorical variables. Test results not within the reference range (Supplement Table 1) were classified as abnormal (i.e., presence of the autoantibody or out of the normal range). Those within the reference range were considered normal. Reference ranges of immunomarkers might be subject to exam method shifts or updates according to years of research findings. In response to the issue, reference range changes over time were considered when converting exam results into dichotomous forms.

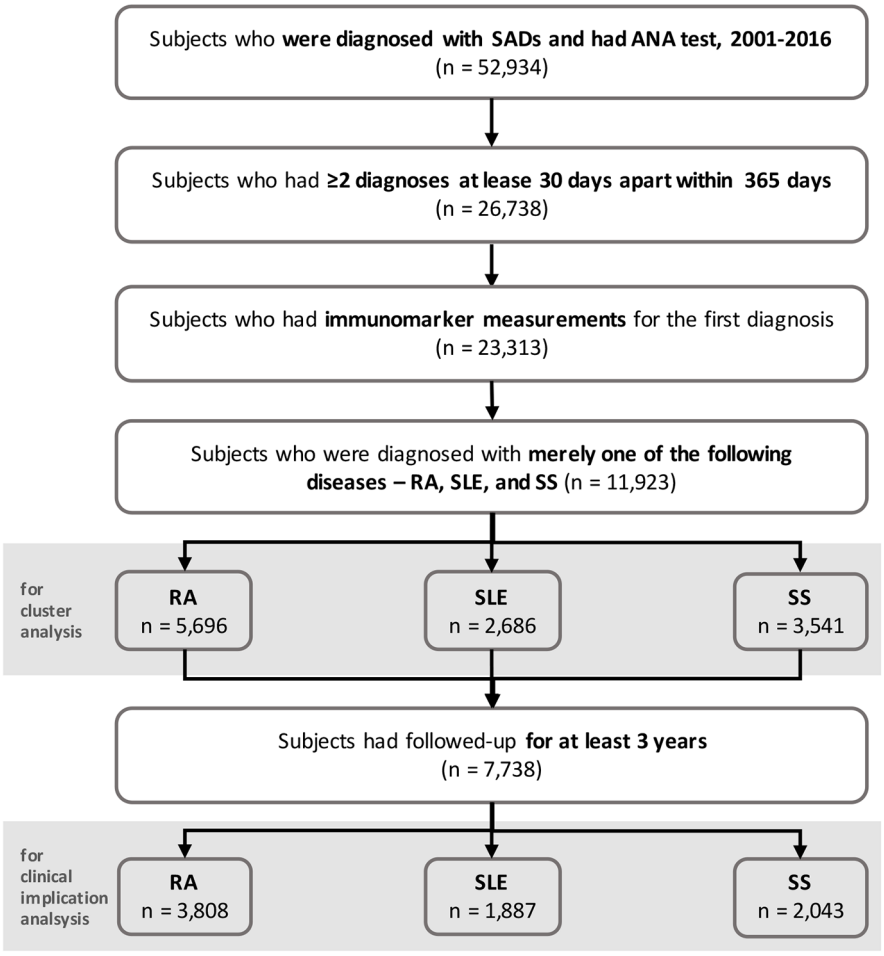

Fig. 1 Flowchart of patient inclusion rules. (RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; SS, Sjögren's syndrome)

The problem of missing values is prominent in a real-world dataset because physicians order different laboratory tests even for the same disease. Generally, we imputed missing values with "normal" results (i.e., negative and nonreactive) except for the rheumatoid factor (RF) and ANA. For diagnosing a specific SAD, laboratory tests that clinical physicians do not order are regarded as minimally valuable, and the test results are assumed to be within the reference range. By contrast, the missing value of the RF was imputed as "abnormal" for patients with RA, given the fact that RF positivity is associated with RA development [22] and is a valuable biomarker for RA diagnosis [23]. The same principle was applied to the missing values of ANA testing for patients with SLE and SS, considering the high diagnostic value of ANA testing in SLE [24] and SS [25]. The missing rates for all the immunomarkers are presented in Supplement Fig. 1.

Cluster analysis

We performed a cluster analysis using MCA k-means. This method joints dimension reduction and cluster analysis by combining MCA with k-means in a unified framework. By contrast, the tandem approach that involves first determining PCs and then clustering patients using these PCs with clustering algorithms has an inherent problem where the first few PCs do not necessarily define the most informative subspace of a genuine

cluster structure of the original dataset. Thus, we employed the joint approach to solve this problem. Details of feature selection and parameter tuning are in Supplement Document 3.

Assessment of clinical implications

To evaluate the performance of the cluster analysis in aggregating patients into clinically significant subgroups, we compared the occurrence of clinical manifestations after SAD diagnosis between groups for patients who were followed for at least 3 years (Fig. 1). The patients who had a follow-up period of of less than 3 years were excluded at this stage to trace the occurrence of occurrence of clinical manifestations after SAD. Clinical manifestations were categorised in two ways: (1) typical or severe manifestations observed in RA, SLE, and SS, defined in Supplement Table 2, and (2) Clinical Classifications Software (CCS) categories, considered to comprehensively screen for clinical manifestations that could be rare and associated with a specific group. Patients' manifestations might be subject to medical treatments; because of the convention that treatments are conditional on diagnoses for the reimbursement purpose of the National Health Insurance (NHI), the temporality of manifestations and treatments is favourable in our study.

Statistical analysis

Analysis of variance (ANOVA) was performed to examine the distribution of differences between the means of SADs. The chi-square test or Fisher's exact test was performed for the univariate analysis of categorical variables. Test for proportion difference of comorbidity, across all clusters and by every two clusters, were conducted by Fisher's exact test and Barnard's unconditional test. Adjusted odd ratios (OR_{adj}) were calculated using a multiple logistic regression model. The R package clustrd [26] was used to implement the MCA k-means algorithm. Statistical analyses were performed using R (version 4.1.0, The R Foundation for Statistical Computing). All statistical tests were two-sided, and a p-value of <0.05 was considered significant for statistical analyses. R scripts for the analyses above are available at https://github.com/DHLab-TSENG/Heterogeneity-in-SAD-paper.

Results

Demographic characteristics and immunomarker

We profiled the data of patients with SS ($n=3,541$), RA ($n=5,696$), and SLE ($n=2,686$) (Fig. 1; Table 1). The patients with SLE (mean age $=33.9\pm17.5$ years) were younger than those with RA (51.0 ± 16.6) and SS (54.2 ± 14.2) ($p<0.001$). More than 85% of the patients with SS and SLE were female, whereas the patients with RA had a slightly lower proportion (74%) in females. The proportion of patients with abnormal immunomarker results varied among the SADs (Table 1).

Clustering results and cluster characteristics

Figure 2A presents the results of the cluster analysis. The best model, with the highest average silhouette width, retained two PCs, resulting in six clusters. The first PC separated Clusters 1, 3, 4, and 6, whereas the second PC further opposed Clusters 2 and 5.

The proportion of patients with the three SADs was imbalanced among the clusters. The patients with RA, SLE, and SS mainly dominated Clusters 1, 3 and 6, and 4 and 5,

Table 1 Demographic characteristics and proportions of immunomarker test results in patients with SS, RA, and SLE (immunomarkers with normal test results for all the SADs were excluded)

Variable	Sjögren's syndrome (*n* = 3,541)	Rheumatoid arthritis (*n* = 5,696)	Systemic lupus erythematosus (*n* = 2,686)	*p*-value
Age, mean (sd)	54·2 (14·2)	51·0 (16·6)	33·9 (17·5)	< 0·001
Sex, n (%)				
Female	3,070 (86·70)	4,250 (74·61)	2,307 (85·89)	< 0·001
Male	471 (13·30)	1,446 (25·39)	379 (14·11)	< 0·001
Autoantibody (= positive), n (%)				
ACAG	12 (0·34)	1 (0·02)	59 (2·20)	< 0·001
ACAM	3 (0·08)	1 (0·02)	3 (0·11)	0·188
ANA	2,049 (57·87)	1,159 (20·35)	2,335 (86·93)	< 0·001
AQP4-autoAb	2 (0·06)	0 (0)	0 (0)	0·094
Anti-BMZA	0 (0)	0 (0)	3 (0·11)	0·006
Anti-CENP	6 (0·17)	0 (0)	0 (0)	< 0·001
Anti-ICSA	0 (0)	1 (0·02)	2 (0·07)	0·164
Anti-Jo	1 (0·03)	0 (0)	0 (0)	0·306
Anti-RNP-Sm	18 (0·51)	8 (0·14)	92 (3·43)	< 0·001
Anti-RibP	0 (0)	0 (0)	2 (0·07)	0·032
Anti-SSA	496 (14·01)	37 (0·65)	155 (5·77)	< 0·001
Anti-SSB	290 (8·19)	18 (0·32)	79 (2·94)	< 0·001
Anti-Scl	3 (0·08)	1 (0·02)	5 (0·19)	0·031
Anti-THYG	28 (0·79)	10 (0·18)	10 (0·37)	< 0·001
Anti-TPO	62 (1·75)	30 (0·53)	34 (1·27)	< 0·001
Anti-TSHR	1 (0·03)	0 (0)	1 (0·04)	0·386
Anti-dsDNA	45 (1·27)	27 (0·47)	1,221 (45·46)	< 0·001
B2GP1G	0 (0)	0 (0)	1 (0·04)	0·179
CRYOFIBRI	83 (2·34)	55 (0·97)	68 (2·53)	< 0·001
CRYOID	215 (6·07)	141 (2·48)	208 (7·74)	< 0·001
DC IgG	1 (0·03)	0 (0)	19 (0·71)	< 0·001
FLC Kappa	5 (0·14)	1 (0·02)	5 (0·19)	0·031
FLC Lambda	4 (0·11)	0 (0)	5 (0·19)	0·009
GAD-Ab	0 (0)	1 (0·02)	0 (0)	0·579
RF	2,258 (63·77)	5,695 (99·98)	817 (30·42)	< 0·001
Complement (= abnormal), n (%)				
C3	239 (6·75)	125 (2·19)	1,513 (56·33)	< 0·001
C4	72 (2·03)	124 (2·18)	1,006 (37·45)	< 0·001

Abbreviations: ACAG = anti-cardiolipin antibody IgG, ACAM = anti-cardiolipin antibody IgM, ANA = anti-nuclear antibody, AQP4-autoAb = anti-aquaporin 4 antibodies, anti-BMZA = basement membrane zone antibody, anti-CENP = anti-centromere antibody, anti-ICSA = intracellular substance antibody, anti-Jo = anti-Jo 1 antibody, anti-RNP-Sm = anti-ribonucleoprotein and anti-Smith antibody, anti-RibP = anti-ribosomal P antibody, anti-SSA = anti-SSA/Ro antibody, anti-SSB = anti-SSB/La antibody, anti-Scl = anti-Scl antibody, anti-THYG = anti-thyroglobulin antibody, anti-TPO = anti-thyroid peroxidase antibody, anti-TSHR = anti-TSH receptor antibody, anti-dsDNA = anti-double stranded DNA antibody, B2GP1G = beta-2 glycoprotein 1 antibody IgG, CRYOFIBRI = cryofibrinogen identification, CRYOID = cryoblobulin identification, DC IgG = direct coombs IgG, FLC Kappa = free light chain kappa, FLC Lambda = free light chain lambda, GAD-Ab = glutamic acid decarboxylase 65 antibody, RF = rheumatoid factor, C3 = complement component C3, C4 = complement component C4

respectively. The patients with SLE and SS were evenly distributed in Cluster 2 (Fig. 2B). Figure 2C depicts the proportion of patients with the SADs in the six clusters. RA accounted for a disproportionately high percentage (>90%) in Cluster 1, whereas SLE and SS had a more dispersed distribution in the six clusters.

We identified a distinctive pattern by determining the proportion of abnormal immunomarker results between the clusters (Fig. 3A and B), and details are described in **Supplementary Document 4**.

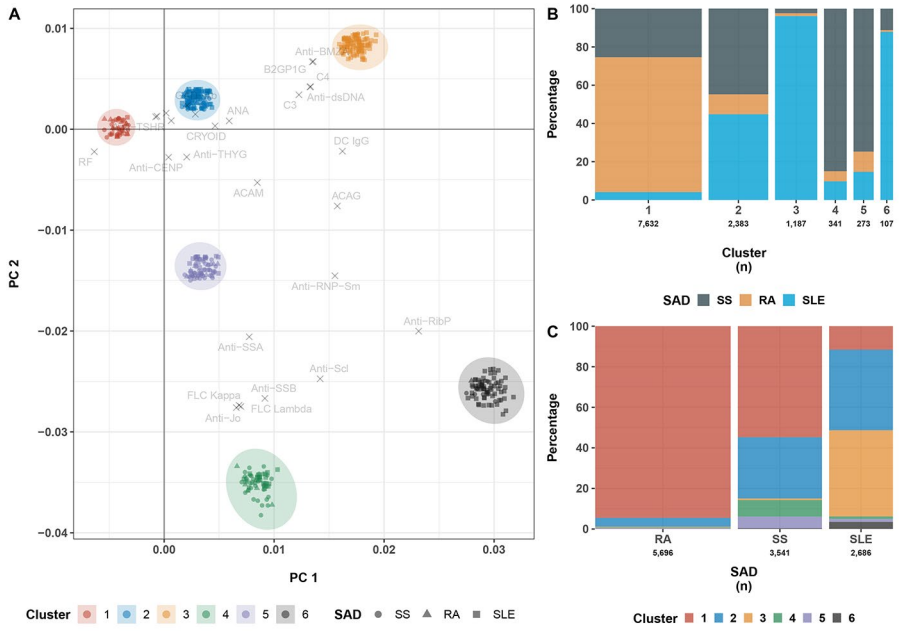

Fig. 2 Diverse clustering of the patients with SADs based on immunomarkers. **(A)** Clusters of patients obtained using MCA k-means with abnormal immunomarkers contributing to the lowest two principal components (PCs). Coloured points with diseases labelled in shapes represent the patients, and immunomarkers with abnormal results are shown as black crosses (x) along with their names. The relative location between a cluster's centroid and an immunomarker suggests the tendency, compared with other clusters, of patients in a specific cluster to test abnormal for immunomarkers located nearby. For example, Cluster 3 is located near C3, C4, and anti-dsDNA, suggesting that patients in this cluster are more likely to test abnormal for these immunomarkers. **(B)** SADs are grouped into six clusters. The patients with the SADs (RA, SLE, and SS) are grouped into six clusters based on the pattern of immunomarkers. Each cluster is composed of multiple SADs. For example, Cluster 2 comprises SLE, SS, and a few RA cases. The heterogeneity indicates that the patients exhibit similar immunomarker patterns even if diagnosed with different SADs. **(C)** Heterogeneity of SADs. RA is predominantly distributed only in Cluster 1. High heterogeneity is noted for SLE and SS – patients with SS can exhibit the immunomarker patterns of Clusters 1, 2, 4, and 5; patients with SLE are also observed in Clusters 2 and 3

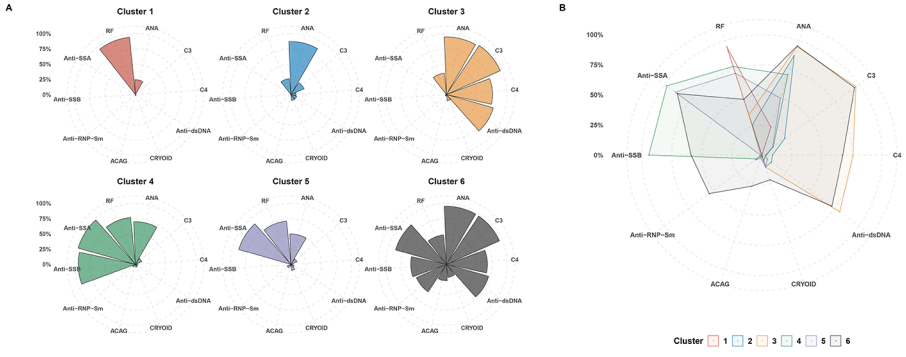

Fig. 3 Immunological characteristics of clusters based on immunomarkers. **(A)** and **(B)** The proportion of abnormal results in the 10 immunomarkers in the clusters

High heterogeneity in clinical manifestations of SADs between the clusters

We determined the rate of clinical manifestations for the SADs grouped into six clusters. SS showed the most heterogeneous clinical manifestations between clusters (Fig. 4A **top**). Patients with SS in Cluster 3 were the most distinctive because of the high occurrence of immunity disorders (63%) and other and unspecified benign neoplasm (58%), compared with the other clusters (all $p<0.05$). More than half of the patients with SS

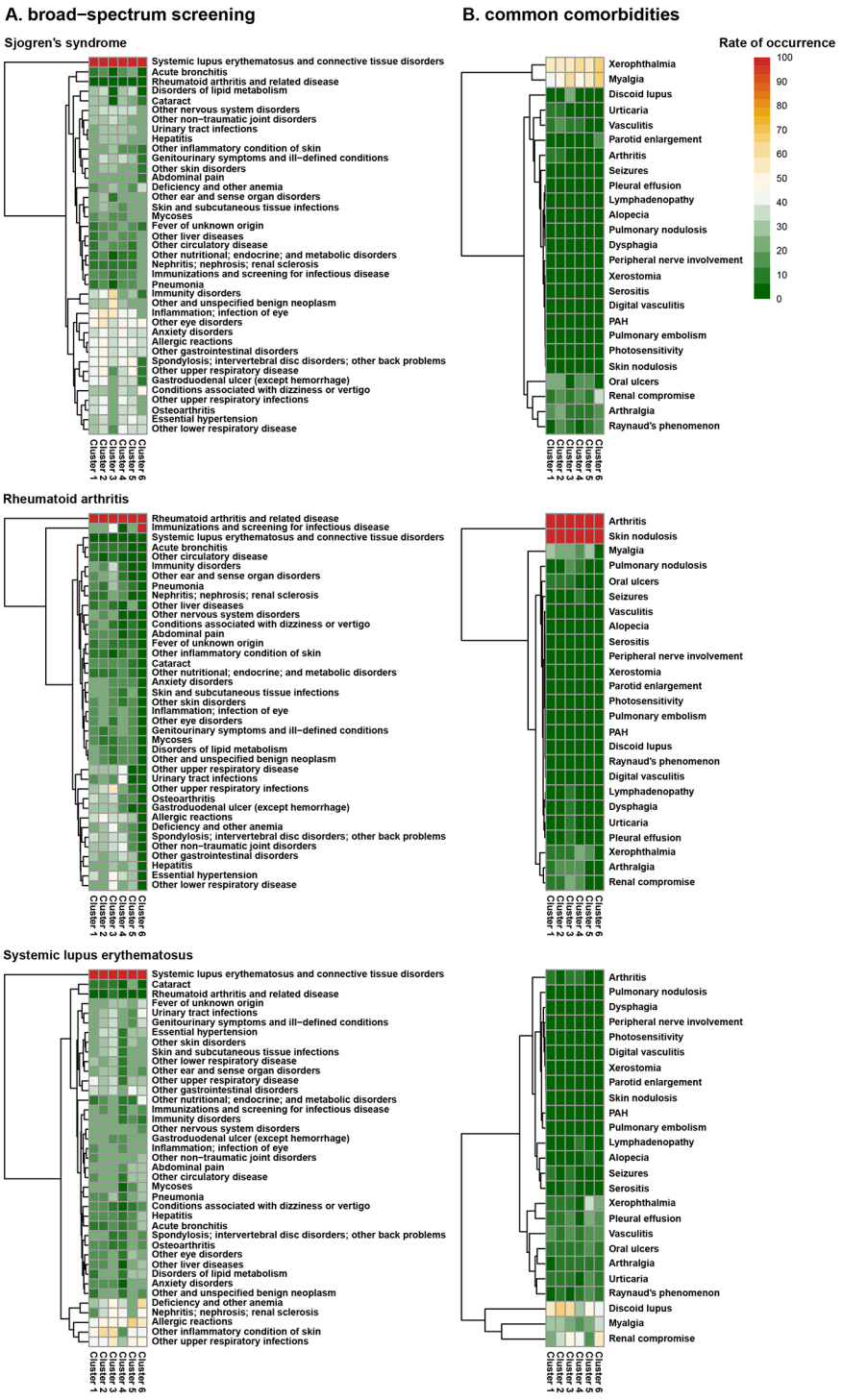

Fig. 4 Heatmaps illustrate the rates of **(A)** Clinical Classification Software (CCS) diagnosis groups and **(B)** common clinical manifestation occurrences of the systemic autoimmune diseases between clusters in each of the individual diseases. We classified all ICD codes (not limited to SADs related codes) into CCS single-level diagnosis groups to identify clinically meaningful manifestations. As shown in Fig. 4B, we collected, from other studies, manifestations that were commonly observed in patients with the SADs

in Clusters 1 (53%), 2 (54%), and 3 (53%) experienced inflammation or infection of the eye; however, Cluster 2 had other symptoms frequently observed in SS higher than the other clusters, such as spondylosis (47%; $p<0.05$) (e.g., backache, lumbago, cervical and lumbosacral spondylosis without myelopathy) and upper respiratory infections (42%; $p<0.04$) compared with Cluster 3 as well as higher upper respiratory disorders (48%) with Cluster 3 ($p=0.01$) and Cluster 4 ($p=0.02$). The patients with SS in Cluster 4 appeared to have less cutaneous symptoms (e.g., inflammatory condition of skin as well as skin and subcutaneous tissue infections) ($p<0.05$) than the other clusters but had a higher risk of manifesting lower respiratory diseases as compared with Cluster 3 ($p=0.03$) (Fig. 4A **top**).

The patients with SLE in different clusters also demonstrated highly diverse clinical manifestations, specifically in blood, renal, and cutaneous-related symptoms (Fig. 4 bottom). Both Clusters 3 and 6 exhibited significantly higher proportions of patients with the two manifestations, where more than half had deficiency and other anemia (50% and 65%, $p<0.001$) and nephritis (52% and 51%, $p<0.001$). Clusters 1, 2, and 3 significantly displayed higher proportions of patients with discoid lupus (55%, 67%, and 60% respectively); however, Cluster 2 had higher proportions of patients with inflammatory conditions of skin (61%) and skin and subcutaneous tissue infections (32%) than Cluster 1 (46%; $p=0.03$ and 23%; $p<0.05$). In contrast, Clusters 4 and 5 manifested less diversity than the other clusters.

Clinical manifestations in RA were relatively homogeneous (Fig. 4A middle). Clusters 1, 2, and 3 had a higher risk of being diagnosed with osteoarthritis than the other three clusters ($p<0.05$). While Cluster 3 had a higher proportion of patients with upper respiratory infections (53%) than those in Cluster 5 (11%, $p=0.01$). Furthermore, Cluster 4 had higher proportions of patients with both upper respiratory disease (42%; $p<0.05$) and urinary tract infections (42%; $p=0.02$) when compared with Cluster 5.

We implemented the same clinical implication analysis but without stratifying by SADs. The heterogeneities in clinical manifestations were still observable (Supplement Fig. 2).

Discussion

We successfully deciphered the heterogeneity in common SADs and obtained six clinically meaningful clusters by applying the simultaneous dimension reduction and cluster analysis approach to immunomarker data. The results revealed relationships between diseases that are not apparent with classical diagnoses alone. Based on the clustering results, parts of SLE and SS cases were clustered together in Cluster 2 (Fig. 2), indicating that these cases, despite having different classical diagnoses, were more closely related at the molecular level. This biomarker-based clustering also showed clinical relevance (Fig. 4), suggesting that re-classification through this approach could provide clinically significant subgroups of SADs. In addition to including common SADs approach, we also clustered specific SADs into subgroups (Fig. 2C). These clusters revealed considerable differences in the abnormality pattern of immunomarkers (Fig. 3), which aligns with the observation in a previous study [19], and the occurrence of clinical manifestations.

Heterogeneity in major SADs

The cluster analysis results revealed heterogeneity in the major SADs (Fig. 2B and C). Moreover, the level of heterogeneity considerably differed among the SADs. High heterogeneity was noted for SS. Patients with SS were majorly clustered into Clusters 1 and 2. Although fewer patients were grouped into Clusters 4 and 5, patients with SS dominated the two clusters. Similar heterogeneity was also identified for SLE. Patients with SLE were predominantly distributed in Clusters 2 and 3, but the proportions of the patients were both higher in Clusters 3 and 6. By contrast, RA revealed relatively low heterogeneity (Fig. 2C).

The pattern of immunomarkers differed among the six clusters. Patients with SLE and SS in Cluster 2 presented more closely than typical cases in Clusters 3 and 6 (typical SLE) and Clusters 4 and 5 (typical SS), respectively. Similarly, the patients with SS in Cluster 1 were similar to those with RA. The results implied that in specific clusters (e.g., Clusters 1 and 2), the SADs would manifest atypically with immune features that can be found in other SADs: patients with SS in Cluster 1 would manifest more RA-like features, whereas SS in Cluster 2 would manifest more SLE-like features. Thus, compared with the current diagnostic standards of SADs, the clustering-based categorization of the SADs would be a more representative and data-driven definition.

Association between immunomarkers and clinical manifestations

A SAD classification system considering heterogeneity, such as in the pattern of immunomarker presentation or manifestations, enables us to delineate various underlying pathological mechanisms for identical diseases classified by the current system. While we reached a similar conclusion with a previous study in the pattern of immunomarkers between clusters [19], we further associated these clusters with clinical manifestations. Our results revealed the links between particular patterns of immunomarker presence (Fig. 3) and the occurrence of clinical manifestations (Fig. 4), providing additional value in identifying SAD subgroups.

Association between immunomarkers and clinical manifestations for SS

A meta-analysis revealed significant associations between primary SS and various malignancies; risks were especially unneglectable in non-Hodgkin lymphoma, leukemia, myeloma, and autoimmune disease malignancy [27]. Moreover, a cohort study revealed significant predictors of lymphoproliferative diseases, such as purpura/skin vasculitis, low C3 levels, and low C4 levels [28], which were also observed in our study (Cluster 3, Fig. 3A).

Xerophthalmia is a common ocular manifestation of SS. Clusters 4 and 6 revealed higher frequencies of dry eye (61% and 67%, respectively) and differences in anti-SSA and anti-SSB results compared to other clusters (Fig. 3A). A study reported that the presence of anti-SSA/Ro concurrently with anti-SSB/La would deteriorate the ocular manifestation of SS [29]. We observed a similar trend where xerophthalmia appeared to be more prevalent in Clusters 4 and 6 (having higher proportions of anti-SSA and anti-SSB positivity) than in Cluster 5 (only anti-SSA positivity).

We identified that Clusters 2, 4, and 5 had more patients with back pain problems (47%, 43%, and 46%, respectively). Although a few studies have reported that patients with SS were affected by spondyloarthritis, the connections between them and the

pathology remained inconclusive [30, 31]. By contrast, Cluster 4 had more patients with lower respiratory tract disease. A previous study revealed that the lower respiratory tract can be affected by SS, including the lung parenchyma, mainly with interstitial lung disease, airways, vasculature, and pleura [32].

Association between immunomarkers and clinical manifestations for SLE

An association between immunomarkers and clinical manifestations was noted in Clusters 3 and 6. Renal compromise, anemia, and inflammatory condition of the skin, which manifests by patients with relapsing or active SLE, are of great concern to prognosis. The patients in Cluster 3 appeared to have a propensity for severe SLE comorbidities. Elevated ds-DNA and decreased C3 and C4 levels were correlated with severe disease activity [33, 34] and renal flares [35]. In addition, progressively simultaneous reductions in C3 and C4 levels and increasing ds-DNA levels may be helpful predictors for future relapses in serologically active clinically quiescent SLE [36].

A cross-sectional study reported that the expression of anti-dsDNA and anti-SSA was higher in patients with SLE without discoid lupus erythematosus (DLE−/SLE+) than in those with DLE (DLE+/SLE+) [37]. We observed a similar trend in our study: Cluster 6, which had lower cutaneous manifestations, had a higher proportion of patients with positive results for anti-SSA than Clusters 1, 2, and 5, which had higher cutaneous manifestations. Moreover, the contribution of anti-RNP-Sm to DLE remains unclear, showing contradictory results in recent studies [37–39].

DLE indicates that patients may undergo a benign lupus course and have a lower risk of renal involvement [37]. However, this trend was only observed in Clusters 1, 2, and 5; we observed a negative correlation – by performing a multiple logistic regression with the covariates of sex, anti-SSA, and anti-RNP-Sm – between discoid lupus and renal compromise for all the patients with SLE in the three clusters (OR_{adj} for discoid lupus = 0·47; 95% CI = 0·34–0·65).

Association between immunomarkers and clinical manifestations for RA

Although no difference in common comorbidities was noted between clusters for RA, several extraarticular manifestations displayed a higher occurrence in some clusters. The most common manifestations were upper respiratory infections and lower respiratory diseases in Cluster 3, and upper respiratory disease and urinary tract infections in Cluster 4. Lung involvement significantly contributes to mortality and morbidity in the extraarticular manifestations of RA. The involvements often correlate with treatment agents, longstanding diseases, infections, and health behaviours [32].

Limitations

The ACR/EULAR 2019 criteria for SLE require ANA positivity. Some SLE diagnoses in our study may not meet these updated criteria. Besides, ANA was chosen as the initial biomarker and patients without ANA test were excluded. Nevertheless, it's important to note that ANA is a nearly essential test for individuals with suspected autoimmune diseases, despite its suboptimal diagnostic accuracy in certain autoimmune conditions like RA. Consequently, while the entry criteria might raise concerns, they are unlikely to introduce bias into the study.

We assumed that the immunomarker results for patients not specifically ordered by a physician were within the reference range, except for RF in RA and ANA in SLE and SS cases. Physicians use expertise to determine test necessity for diagnosis and tailor treatments in clinical settings. This assumption, while introducing bias, seemed plausible because SAD symptoms strongly correlate with autoantibody presence, helping us address missing data mechanisms. Additionally, the plausibility and efficacy of imputing missing values with normal values have been reported [15]. Notably, ANA seronegativity in SLE cases is rare, and post-diagnostic ANA seropositivity is common in primary SS [16, 17]. Therefore, we imputed RF and ANA antibodies in a disease-specific manner. Moreover, some diagnostically valuable tests not covered by NHI (e.g., anti-citrullinated peptide antibody) were excluded due to high missing rates.

Other limitations include coding bias when using the diagnosis codes to identify clinical manifestations. We used an inclusion algorithm (codes appearing at least twice within a specified time) to reduce misclassification. Major SADs are not limited to RA, SLE, and SS, and patient with multiple SADs is not uncommon – constituting nearly a quarter (24.9%) of SAD patients in our database. While the prevalence of certain other SADs, like systemic sclerosis (SSc), is relatively lower in our database, we intend to continue collecting data on patients with various SADs and those with multiple SADs for future in-depth analyses.

Finally, our study primarily focused on Han Taiwanese patients, which limits the generalizability to other ethnicities.

Conclusion

We developed an immunomarker–driven clustering method for major SADs based on large-scale data. High heterogeneity was identified for the conventional SADs among these immunomarker-based clusters. Moreover, clinical manifestations correlated well with the immunomarker-driven clustering, implying that the clustering method could recognise more clinically relevant subgroups. By identifying these clinically relevant subgroups, physicians can focus on specific risks and tailor their management strategies accordingly. This approach facilitates personalized treatment plans and ensures that high-related features are closely monitored.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00389-7.

Supplementary Material 1

Author contributions

C-W. C and Y-J. T had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. C-W. C analysed/interpreted the data and performed experiments. C-W. C, H-Y. W, and Y-J. T designed the study and wrote the paper. W-L. L, W-T. L, and J-R. Y reviewed/edited the manuscript for important intellectual content and provided administrative, technical, or material support. Y-J. T obtained funding and supervised the study.

Funding

This research was supported by the National Science and Technology Council, Taiwan under Grants NSTC 111-2320-B-182 A-002-MY2 and 111-2628-E-A49-026-MY3; Chang Gung Memorial Hospital under Grant CORPG2P0071; the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan; Keelung Chang Gung Memorial Hospital and National Yang Ming Chiao Tung University Joint Research Program under Grants CGMH-NYCU-113-CORPG2P0071.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Chang Gung Medical Foundation (202001298B0), and the requirement for patient consent was waived.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2]Department of Laboratory Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan
[3]Syu Kang Sport Clinic, Taipei, Taiwan
[4]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
[5]Department of Rheumatology, Chang Gung Memorial Hospital at Keelung, Keelung, Taiwan
[6]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
[7]School of Medicine, National Tsing Hua University, Hsinchu, Taiwan
[8]Department of Medicine, University of California San Diego, San Diego, CA, USA
[9]John A. Burns School of Medicine, University of Hawaii, Honolulu, HI, USA
[10]Queen's Heart Institute, Queens Medical Center, Honolulu, HI, USA

Received: 23 April 2024 / Accepted: 2 September 2024
Published online: 16 September 2024

References

1. Cojocaru M, Cojocaru IM, Silosi I, Vrabie CD. Manifestations of systemic lupus erythematosus. Maedica [Internet]. 2011;6:330–6. Available from: /pmc/articles/PMC3391953/.
2. Arnaud L, Tektonidou MG. Long-term outcomes in systemic lupus erythematosus: trends over time and major contributors. Rheumatology [Internet]. 2020;59:v29–38. https://academic.oup.com/rheumatology/article/59/Supplement_5/v29/6024730?login=true
3. Mosca M, Tani C, Talarico R, Bombardieri S. Undifferentiated connective tissue diseases (UCTD): Simplified systemic autoimmune diseases. Autoimmun Rev [Internet]. 2011;10:256–8. https://www.sciencedirect.com/science/article/pii/S1568997210002090?via%3Dihub
4. Committee ADAPP. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022. Diabetes Care [Internet]. 2022;45:S17–38. https://diabetesjournals.org/care/article/45/Supplement_1/S17/138925/2-Classification-and-Diagnosis-of-Diabetes
5. Thygesen K, Alpert JS, Jaffe AS, Simoons ML, Chaitman BR, White HD et al. Third universal definition of myocardial infarction. Eur Heart J [Internet]. Eur Heart J; 2012. pp. 2551–67. https://academic.oup.com/eurheartj/article/33/20/2551/447556
6. Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA et al. Fourth universal definition of myocardial infarction (2018). Eur Heart J [Internet]. 2019;40:237–69. https://www.ahajournals.org/doi/https://doi.org/10.1161/CIR.0000000000000617
7. Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. Ann Rheum Dis [Internet]. 2019;78:1151. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=31383717&retmode=ref&cmd=prlinks
8. Fenton KA, Pedersen HL. Advanced methods and novel biomarkers in autoimmune diseases a review of the recent years progress in systemic lupus erythematosus. Front Med. 2023;10:1183535.
9. Borrebaeck CAK. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. Nat Rev Cancer. 2017;17:199–204.
10. Wen Y-H, Chang P-Y, Hsu C-M, Wang H-Y, Chiu C-T, Lu J-J. Cancer screening through a multi-analyte serum biomarker panel during health check-up examinations: results from a 12-year experience. Clin Chim Acta. 2015;450:273–6.
11. Wang H-Y, Chen C-H, Shi S, Chung C-R, Wen Y-H, Wu M-H, et al. Improving Multi-tumor Biomarker Health Check-Up tests with machine learning algorithms. Cancers. 2020;12:1442.
12. Wu X, Wang H-Y, Shi P, Sun R, Wang X, Luo Z, et al. Long short-term memory model – a deep learning approach for medical data with irregularity in cancer predication with tumor markers. Comput Biol Med. 2022;144:105362.
13. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. Sci Rep-uk [Internet]. 2017;7:2427. https://www.nature.com/articles/s41598-017-02606-2
14. Orange DE, Agius P, DiCarlo EF, Robine N, Geiger H, Szymonifka J et al. Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. Arthritis Rheumatol [Internet]. 2018;70:690–701. https://doi.org/10.1002/art.40428
15. Barturen G, Babaei S, Català-Moll F, Martínez-Bueno M, Makowska Z, Martorell-Marugán J et al. Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. Arthritis Rheumatol [Internet]. 2021;73:1073–85. https://onlinelibrary.wiley.com/doi/full/10.1002/art.41610
16. Moritz CP, Paul S, Stoevesandt O, Tholance Y, Camdessanché J-P, Antoine J-C, Autoantigenomics. Holistic characterization of autoantigen repertoires for a better understanding of autoimmune diseases. Autoimmun Rev [Internet]. 2020;19:102450. https://www.sciencedirect.com/science/article/pii/S1568997219302630?via%3Dihub
17. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. Npj Digital Medicine [Internet]. 2020;3:30. https://www.nature.com/articles/s41746-020-0229-3

18. Spielmann L, Nespola B, Séverac F, Andres E, Kessler R, Guffroy A et al. Anti-Ku syndrome with elevated CK and anti-Ku syndrome with anti-dsDNA are two distinct entities with different outcomes. Ann Rheum Dis [Internet]. 2019;78:1101. http://ard.bmj.com/lookup/doi/https://doi.org/10.1136/annrheumdis-2018-214439
19. Molano-González N, Rojas M, Monsalve DM, Pacheco Y, Acosta-Ampudia Y, Rodríguez Y et al. Cluster analysis of autoimmune rheumatic diseases based on autoantibodies. New insights for polyautoimmunity. J Autoimmun [Internet]. 2019;98:24–32. https://www.sciencedirect.com/science/article/pii/S0896841118305328?via%3Dihub
20. Shao S, Chan Y, Yang YK, Lin S, Hung M, Chien R et al. The Chang Gung Research Database—A multi-institutional electronic medical records database for real-world epidemiological studies in Taiwan. Pharmacoepidem Drug Safe [Internet]. 2019;28:593–600. https://onlinelibrary.wiley.com/doi/full/https://doi.org/10.1002/pds.4713
21. Tseng Y-J, Chiu H-J, Chen CJ. dxpr: an R package for generating analysis-ready data from electronic health records—diagnoses and procedures. Peerj Comput Sci [Internet]. 2021;7:e520. https://peerj.com/articles/cs-520
22. Kuriya B, Cheng CK, Chen HM, Bykerk VP. Validation of a prediction rule for development of rheumatoid arthritis in patients with early undifferentiated arthritis. Ann Rheum Dis [Internet]. 2009;68:1482. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19015211&retmode=ref&cmd=prlinks
23. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO et al. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheumatism [Internet]. 2010;62:2569–81. https://onlinelibrary.wiley.com/doi/full/https://doi.org/10.1002/art.27584
24. Mutasim DF, Adams BB. A practical guide for serologic evaluation of autoimmune connective tissue diseases. J Am Acad Dermatol [Internet]. 2000;42:159–76. https://www.jaad.org/article/S0190-9622(00)70098-3/fulltext
25. Veenbergen S, Kozmar A, van Daele PLA, Schreurs MWJ. Autoantibodies in Sjögren's syndrome and its classification criteria. J Transl Autoimmun [Internet]. 2022;5:100138. https://www.sciencedirect.com/science/article/pii/S2589909021000587
26. Markos A, D'Enza AI, van de Velden M. Beyond Tandem Analysis: Joint Dimension Reduction and Clustering in R. J Stat Softw [Internet]. 2019;91. https://www.jstatsoft.org/index.php/jss/article/view/v091i10/v91i10.pdf
27. Liang Y, Yang Z, Qin B, Zhong R. Primary Sjögren's syndrome and malignancy risk: a systematic review and meta-analysis. Ann Rheum Dis [Internet]. 2014;73:1151. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23687261&retmode=ref&cmd=prlinks
28. Theander E, Henriksson G, Ljungberg O, Mandl T, Manthorpe R, Jacobsson LTH. Lymphoma and other malignancies in primary Sjögren's syndrome: a cohort study on cancer incidence and lymphoma predictors. Ann Rheum Dis [Internet]. 2006;65:796. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=16284097&retmode=ref&cmd=prlinks
29. Baer AN, DeMarco MM, Shiboski SC, Lam MY, Challacombe S, Daniels TE et al. The SSB-positive/SSA-negative antibody profile is not associated with key phenotypic features of Sjögren's syndrome. Ann Rheum Dis [Internet]. 2015;74:1557. http://pubmed.gov/25735642
30. Jarrot P-A, Arcani R, Darmon O, Roudier J, Cauchois R, Mazodier K et al. Axial Articular Manifestations in Primary Sjögren Syndrome: Association With Spondyloarthritis. J Rheumatology [Internet]. 2020;48:1037–46. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=32669446&retmode=ref&cmd=prlinks
31. Eren R, Can M, Alibaz-Öner F, Yilmaz-Oner S, Yilmazer B, Cefle A et al. Prevalence of inflammatory back pain and radiologic sacroiliitis is increased in patients with primary Sjögren's syndrome. Pan Afr Medical J [Internet]. 2018;30:98. https://www.panafrican-med-journal.com/content/article/30/98/full/
32. Luppi F, Sebastiani M, Sverzellati N, Cavazza A, Salvarani C, Manfredi A. Lung complications of Sjogren syndrome. European Respir Rev [Internet]. 2020;29:200021. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=32817113&retmode=ref&cmd=prlinks
33. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH, Austin A et al. Derivation of the sledai. A disease activity index for lupus patients. Arthritis Rheumatism [Internet]. 1992;35:630–40. https://onlinelibrary.wiley.com/doi/full/10.1002/art.1780350606
34. Gladman DD, Ibañez D, Urowitz MB. Systemic lupus erythematosus disease activity index 2000. J Rheumatol [Internet]. 2002;29:288–91. Available from:. jrheum.org/content/29/2/288.long.
35. Narayanan K, Marwaha V, Shanmuganandan K, Shankar S. Correlation between systemic lupus erythematosus disease activity index, C3, C4 and anti-dsDNA antibodies. Medical J Armed Forces India [Internet]. 2010;66:102–7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4920905/
36. Miyawaki Y, Sada K, Asano Y, Hayashi K, Yamamura Y, Hiramatsu S et al. Progressive reduction of serum complement levels: a risk factor for relapse in patients with hypocomplementemia in systemic lupus erythematosus. Lupus [Internet]. 2018;27:2093–100. https://journals.sagepub.com/doi/10.1177/0961203318804892?url_ver=Z39.88-2003_id=ori:rid:crossref.org_dat=cr_pub%20%200pubmed
37. Chong BF, Tseng L, Lee T, Vasquez R, Li QZ, Zhang S et al. IgG and IgM Autoantibody Differences in Discoid and Systemic Lupus Patients. J Invest Dermatol [Internet]. 2012;132:2770–9. https://www.sciencedirect.com/science/article/pii/S0022202X1535538X
38. Vasquez R, Tseng L, Victor S, Zhang S, Chong BF. Autoantibody and Clinical Profiles in Patients With Discoid Lupus and Borderline Systemic Lupus. Arch Dermatol [Internet]. 2012;148:651–5. https://jamanetwork.com/journals/jamadermatology/fullarticle/1160677
39. Kim A, O'Brien J, Tseng L-C, Zhang S, Chong BF. Autoantibodies and Disease Activity in Patients With Discoid Lupus Erythematosus. Jama Dermatol [Internet]. 2014;150:651–4. https://jamanetwork.com/journals/jamadermatology/fullarticle/1862049

Publisher's note