Clinical phenotype with high risk for initiation of biologic therapy in rheumatoid arthritis: a data-driven cluster analysis

S.M. Jung, K.-S. Park, K.-J. Kim

Division of Rheumatology, Department of Internal Medicine, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea.

Abstract

Objective

The clinical manifestations and treatment outcome in patients with rheumatoid arthritis (RA) are heterogeneous. We classified RA patients into subgroups with distinct phenotypes through unsupervised clustering and evaluated the utility of this subclassification for evaluation of clinical outcome.

Methods

A total of 1,103 patients with RA were clustered in an unbiased manner using a k-means clustering method, based on their clinical and phenotypic profiles. Initiation of biological disease-modifying anti-rheumatic drugs (bDMARDs) was evaluated in the segregated clusters to investigate the differential clinical course of each cluster.

Results

Patients with RA were classified into four clusters, each with distinct phenotypes. The key features for subclassification were sex, smoking, hypertension, and dyslipidaemia. Cluster 1 consisted of male smokers, who were most likely to initiate bDMARDs by 30 months (p=0.04). Multivariate analysis revealed that overweight, smoking, erythrocyte sedimentation rate, autoantibodies of high titre, and disease activity were the independent predictors of bDMARD initiation at 30 months. Cluster 1 was the highest or the second highest for these independent predictors, suggesting that cluster 1 contained a high-risk group for early initiation of bDMARDs.

Conclusion

The unsupervised clustering of RA patients demonstrated the feasibility of the novel subclassification with respect to predicting clinical outcome. Identifying high-risk patients by a combination of clinical parameters may be useful for the management of RA.

Key words

rheumatoid arthritis, unsupervised clustering, biologic DMARDs

Seung Min Jung, MD, PhD
Kyung-Su Park, MD, PhD
Ki-Jo Kim, MD, PhD

Please address correspondence to:
Ki-Jo Kim,
Division of Rheumatology,
Department of Internal Medicine,
St. Vincent's Hospital,
93 Jungbu-daero,
Paldal-gu, Suwon,
Gyeonggi-do 16247, Republic of Korea
E-mail: md21c@catholic.ac.kr

Received on June 24, 2020; accepted in revised form on October 12, 2020.

Competing interests: none declared.

Introduction

Rheumatoid arthritis (RA) is a chronic autoimmune inflammatory disease characterised by synovial joint inflammation and hyperplasia, autoantibody production, and joint destruction, which may lead to structural and functional joint impairments and an associated decrease in quality of life (1). RA is a complex and heterogeneous disease; the development and formation of autoantibodies is affected by interactions between multiple genetic and environmental factors (2). Distinct cellular and molecular patterns have been identified from synovial tissue samples of patients with RA (3), which show variable clinical responses to different treatments with only a subset gaining clinical remission or reduced disease activity (4). Disease complexity and heterogeneity are not adequately translated into current clinical subclassification, *i.e.* female *versus* male, seropositive *versus* seronegative, young-onset versus elderly-onset, and early *versus* established stages.

Recent advances in cluster analysis have successfully tackled multidimensional heterogeneous clinical data (5-9). In some inflammatory diseases such as asthma, Crohn's disease, and chronic graft-*versus*-host disease, a single disease entity has been successfully stratified into several phenotypes using unsupervised machine learning methods, and subgroups with different phenotypes have shown different clinical outcomes or molecular backgrounds (7, 8, 10). In a similar fashion, we hypothesise that applying cluster analysis to clinical phenotyping will identify novel patterns in multidimensional data obtained from patients with RA. We further hypothesise that the identified subgroups of patients with RA will have distinct clinical profiles and differential clinical disease progression. We therefore investigated the utility of the clusters by analysing patients with RA.

Methods

Patients

A total of 1,433 RA patients who fulfilled the 2010 RA classification criteria (11) and received care at St. Vincent's Hospital, the Catholic University of Korea (Suwon, Republic of Korea) between 2003 and 2018 were identified. Clinical and laboratory data, radiographic images, and drug prescriptions were retrieved from patient medical records. From these, 1,103 patients who were first diagnosed with RA for onset of arthritic symptoms within 1 year were identified. The study subjects were maximally followed up to 194 months. Biologic disease-modifying anti-rheumatic drugs (bDMARDs) were counted if they were used during the entire follow-up period. Conventional DMARDs (cDMARDs) included methotrexate, hydroxychloroquine, sulfasalazine, leflunomide, and tacrolimus. bDMARDs included tumour necrosis factor inhibitors (etanercept, adalimumab, infliximab, golimumab), tocilizumab, abatacept, rituximab, and tofacitinib. Disease activity was assessed using the Disease Activity Score in 28 joints (DAS28) using C-reactive protein (CRP) level (12). The study was carried out in accordance with the Helsinki Declaration and approved by the institutional review board of St. Vincent's Hospital, the Catholic University of Korea (no. VC19RISI0255). Since this is a retrospective study, informed patient consent was waived.

Assay of RA-associated antibodies

Anti-citrullinated protein antibody (ACPA) was analysed by chemiluminescent microparticle immunoassay (Abbott Laboratories, IL, USA) and a positive reading was defined with a cut-off value of 5 U/mL. Maximum antibody concentration was defined as 200 U/mL, and for statistical purposes, a value of 200 U/mL was assigned to all measurements >200 U/mL. ACPA was divided into three categories: <5 U/mL (negative), 5–200 U/mL (low to moderate level), and >200 U/mL (high level) in line with a previous study (13). Rheumatoid factor (RF) IgM titres were measured with a latex agglutination test (Beckman Coulter, CA, USA) with a cut-off value of 14 U/mL. RF levels were also divided into three categories: <14 U/mL (negative), 14–100 U/mL (low to moderate level) and >100 U/mL (high level) in line with a previous study (14).

Radiographic evaluation

Anteroposterior radiographs of the

hands were scored by two experienced readers using the van der Heijde modified Sharp score (SHS) (15). The films were scored in chronological order, and the readers were blinded to all patient data. The potential maximum total score for both hands is 280 (16 areas scored for erosions [score 0–5] and 15 areas for joint space narrowing [score 0–4] in each hand). The interobserver reliability was assessed by calculating the intraclass correlation coefficient which was 0.861 (95% confidence interval [CI], 0.779–0.922).

Cluster analysis

Clinical, laboratory, and radiographic variables characterising RA (sex, age, body mass index (BMI), diabetes, hypertension, dyslipidaemia, smoking, erythrocyte sedimentation rate (ESR), CRP, DAS28, RF, ACPA, SHS, osteoporosis) were broken up into clusters using a *k*-means clustering method. *k*-means clustering aims to partition *n* observations into *k* clusters, in which each observation belongs to the cluster with the nearest mean, or centroid, which serves as the prototype for the cluster (16). Once the location of each centroid is known, future data can be classified by comparing the location of each new data point to the location of each cluster centroid. An inherent risk of the clustering process is the presence of large numbers of samples near the cluster boundaries, which can lead to changes in clustered groups if the analysis is run multiple times. Therefore, despite iterating each individual *k*-means algorithm multiple times with randomly seeded initial centroid locations, convergence to a global optimum is not guaranteed. To accommodate this, the *k*-means clustering algorithm (each with 100 iterations) was repeated 1000 times and the model with the highest number of common iterations (for each given assignment of *k*) was selected. To interpret the strength of each clustering output for multiple assignments of *k*, the within-cluster sum of squared errors (WSS) was examined (17). The squared error for each point is the square of the distance of the point from its representation and the WSS score is the sum of these squared errors for all the points. Optimal *k* is identified as when the WSS first starts to diminish (the 'elbow' method). The importance of features was weighted in order of the feature which minimises within-cluster distance and maximises between-cluster distance using Gower's distance (18, 19).

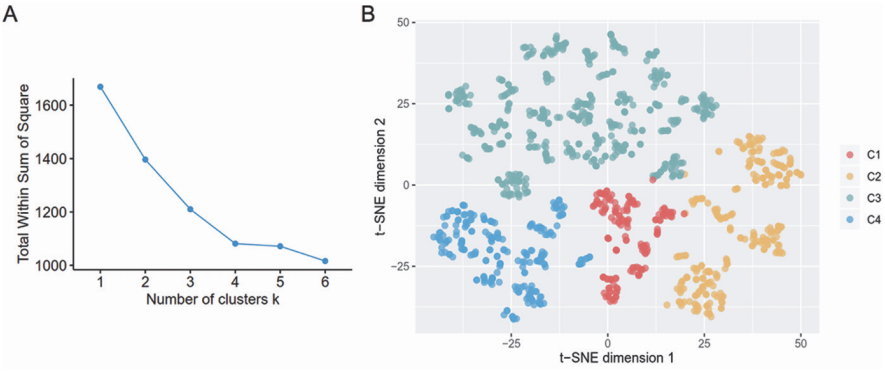

Fig. 1. Unsupervised clustering in the RA cohort by *k*-means cluster analysis.
A: Within-cluster sum of squared errors (WSS) were examined for the *k* from 2 to 6. *k* is optimal at 4 as the WSS starts to diminish (the 'elbow' method).
B: Two-dimensional projection of *t*-SNE result for the RA patients' data. Each patient was coloured by the assigned clusters.

To confirm unsupervised clustering results, we used *t*-distributed stochastic neighbourhood embedding (*t*-SNE), a powerful dimensionality reduction method (20). The *t*-SNE method captures the variance in the data by attempting to preserve the distances between data points from high to low dimensions without any prior assumptions about the data distribution.

Statistical analyses

For deriving the cluster prediction tree, Classification and Regression Tree (CART) analysis using the R package rpart, based on the Gini impurity index, was used (21, 22). The CART model partitioned the data and assigned a predicted class to each subgroup. With repetition of the same process on each predictor in the model, CART identified the best overall split by iteratively testing all possible splits and creating a specified number of nodes, until the specified stopping criteria were reached or a further reduction in node impurity became impossible (23). The partitioning in CART can be represented graphically as an easily interpretable decision tree that may then be used to inform clinical practice (21).

For continuous distributed data, the results are shown as means with standard deviation or medians with interquartile ranges (IQRs); between-group comparisons were performed using a Student's *t*-test or Mann-Whitney U-test. Categorical or dichotomous variables are expressed as frequencies with percentages and were compared using the chi-squared test or Fisher's exact test. Initiation of bDMARDs with corresponding 95% CIs were estimated by Kaplan-Meier analysis and compared using log-rank tests. To identify significant predictors of bDMARD initiation, clinically relevant variables were entered into a multivariable Cox proportional hazards regression model. A two-sided *p*-value of less than 0.05 was considered statistically significant. All statistical analyses were performed using *R* (v. 3.6.1, The *R* Project for Statistical Computing, www.r-project.org).

Results

Clustering of the study population and their characteristics

Clinical, laboratory, and radiographic variables characterising RA (sex, age, BMI, diabetes, hypertension, dyslipidaemia, smoking, ESR, CRP, DAS28, RF, ACPA, SHS, bone mineral density) from 1,103 patients were entered into a *k*-means clustering model. To identify the optimal number of clusters and assess robustness of the clustering, we computed the WSS for different numbers of clusters from 2 to 6 and found that a sum of 4 clusters could optimally represent our data (Fig. 1A). The RA clusters were labelled as C1, C2, C3, and C4. Segregation of the clusters was also reproduced by *t*-SNE, which is an

Table I. Baseline characteristics of the study subjects classified by cluster.

	Cluster				
	C1 (n=145)	C2 (n=256)	C3 (n=479)	C4 (n=223)	*p*-value
Female, n (%)	1 (0.7)	238 (93.0)	440 (91.9)	206 (92.4)	<0.001
Age at diagnosis, years	61 [50, 68]	60 [54, 69]	49 [39, 57]	53 [48, 60]	<0.001
BMI, n (%)					<0.001
Underweight	7 (4.8)	8 (3.1)	52 (10.9)	13 (5.8)	
Normal	97 (66.9)	147 (57.4)	353 (73.7)	151 (67.7)	
Overweight	37 (25.5)	78 (30.5)	59 (12.3)	47 (21.1)	
Obese	4 (2.8)	23 (9.0)	15 (3.1)	12 (5.4)	
Smoking, n (%) †	145 (100.0)	19 (7.4)	33 (6.9)	11 (4.9)	<0.001
Diabetes, n (%)	28 (19.3)	64 (25.0)	13 (2.7)	17 (7.6)	<0.001
Hypertension, n (%)	48 (33.1)	256 (100.0)	0 (0.0)	0 (0.0)	<0.001
Dyslipidaemia, n (%)	58 (40.0)	136 (53.1)	0 (0.0)	223 (100.0)	<0.001
ESR, mm/h	42 [30, 72]	49 [29, 73]	40 [24, 61]	40 [29, 64]	0.010
CRP, mg/dL	1.2 [0.4, 3.9]	0.7 [0.2, 2.3]	0.5 [0.1, 1.9]	0.5 [0.2, 1.7]	<0.001
IgM RF					
Positive, n (%)	114 (78.6)	211 (82.4)	394 (82.3)	200 (89.7)	0.025
Titre, IU/mL	100.0 [23.9, 294.0]	53.6 [24.1, 133.8]	58.2 [21.5, 149.3]	70.7 [30.1, 176.8]	0.006
Subgroup, n (%)			<0.001		
Negative	31 (21.4)	45 (17.6)	85 (17.7)	23 (10.3)	
Low to moderate	41 (28.3)	128 (50.0)	227 (47.4)	111 (49.8)	
High	73 (50.3)	83 (32.4)	167 (34.9)	89 (39.9)	
ACPA					
Positive, n (%)	122 (84.1)	214 (83.6)	403 (84.1)	198 (88.8)	0.353
Titre, IU/mL	91.1 [20,0, 200.0]	81.8 [17.6, 176.2]	92.4 [15.6, 200.0]	86.0 [26.2, 173.7]	0.810
Subgroup, n (%)			0.046		
Negative	23 (15.9)	42 (16.4)	76 (15.9)	25 (11.2)	
Low to moderate	77 (53.1)	159 (62.1)	265 (55.5)	148 (66.4)	
High	45 (31.0)	55 (21.5)	137 (28.6)	50 (22.4)	
SHS, units	0.0 [0.0, 1.0]	0.0 [0.0, 2.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.057
DAS28, units	4.0 [3.4, 4.5]	3.9 [3.4, 4.4]	3.9 [3.4, 4.3]	3.9 [3.3, 4.3]	0.568
Bone mineral density					
L-spine					0.001
Normal	46 (31.7)	59 (23.0)	179 (37.4)	82 (36.8)	
Osteopenia	99 (68.3)	197 (77.0)	300 (62.6)	141 (63.2)	
Osteoporosis	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	
Femur					<0.001
Normal	46 (31.7)	74 (28.9)	189 (39.5)	101 (45.3)	
Osteopenia	78 (53.8)	134 (52.3)	242 (50.5)	104 (46.6)	
Osteoporosis	21 (14.5)	48 (18.8)	48 (10.0)	18 (8.1)	

ACPA: anti-cyclic citrullinated protein antibody; BMI: body mass index; CRP: C-reactive protein; DMARD: disease-modifying anti-rheumatic disease; ESR; erythrocyte sedimentation rate; RF: rheumatoid arthritis; SHS: van der Heijde modified Sharp score.
†Include ex- and current smokers.
§Biologic DMARDs were counted if they were ever used during the whole follow-up period. Conventional DMARDs include methotrexate, hydroxychloroquine, sulfasalazine, leflunomide, and tacrolimus. Biologic DMARDs include tumour necrosis factor inhibitors (etanercept, adalimumab, infliximab, golimumab), tocilizumab, abatacept, rituximab, and tofacitinib.

unsupervised machine learning algorithm that projects all patients onto a two-dimensional plane by reducing dimensionality (Fig. 1B).

The characteristics of the clusters are compared in Table I and Fig. 2A. Members of the C1 subgroup were almost all male smokers and median CRP level and the proportion of patients with high titre of RF and ACPA were highest in this group. The C2 subgroup mostly consisted of hypertensive females and had the highest rate of diabetes and overweight/obesity. Median ESR level and the proportion of patients with osteoporosis were also highest in this group. C3 was the largest subgroup and included mostly younger females without hypertension or dyslipidaemia. The C4 subgroup was mainly composed of non-hypertensive females with dyslipidaemia and had the lowest rate of osteoporosis. With respect to age at diagnosis, members of C1 and C2 were older than those of C3 and C4. Patients with erosive change at baseline were highest in C2 and lowest in C3. However, DAS28 at baseline was not significantly different across the clusters. To obtain a key combination of clinical factors that was predictive of the clusters, a decision tree model was constructed using the CART algorithm (Fig. 2B). Sex, hypertension, dyslipidaemia, and smoking status were four key determining factors to split the clusters, and the patients were perfectly classified in four steps. The four features were confirmed by the importance which minimises within-cluster distance and maximises between-cluster distance as measured by Gower's distance (18, 19).

Risk of initiation of biologic disease-modifying anti-rheumatic drugs

Methotrexate (MTX)-based cDMARD combinations constitute the initial therapy for RA and bDMARDs are administered only if disease activity is not sufficiently controlled despite the use

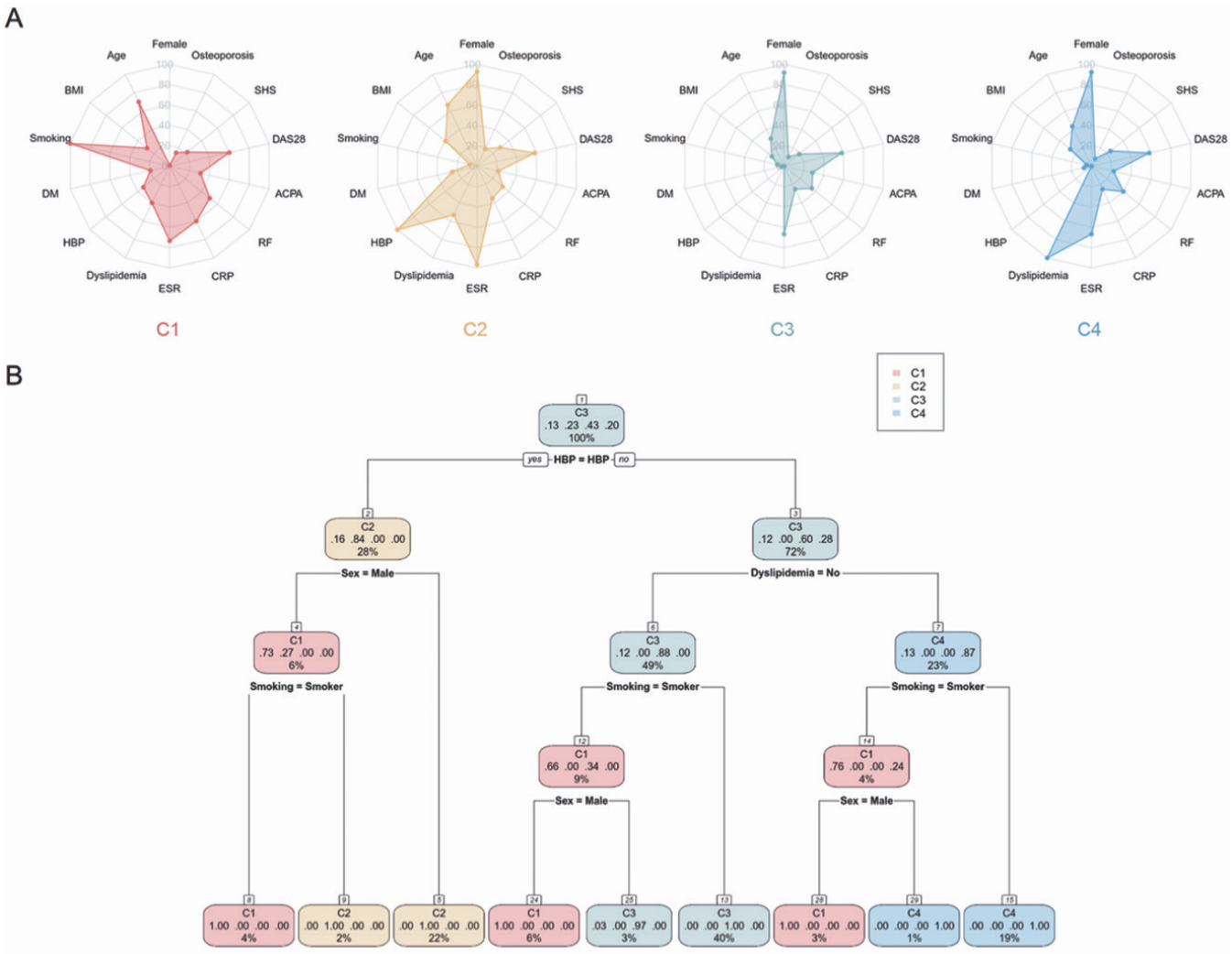

Fig. 2. Clinical patterns and classification of the assigned cluster.
A: Radar chart representing the clinical patterns for the assigned cluster. Each spoke of the radar chart represents a RA-characterising parameter used for the *k*-means cluster analysis. Female, smoking, DM, HBP, dyslipidaemia, and osteoporosis were expressed as percentages. BMI was expressed as the percentage of overweight and obese individuals. Age at diagnosis, ESR, CRP, and DAS28 were relatively expressed for the mean values. SHS was expressed as the percentage of the patients with erosive change. RF and ACPA were expressed as the percentage of patients with high titre.
B: Decision tree by classification and regression trees (CART) analysis. Each node shows the predicted class (C1, C2, C3, or C4), the predicted probability of each class, and the percentage of observation in line order. If the condition is true, go left and down, if not, go right and down.

of MTX-based double or triple combinations of cDMARDs over 6 months according to the National Health Insurance regulations in South Korea. Lack of early response to treatment and initiation of bDMARDs is a surrogate index of poor long-term outcome (24). Cumulative incidence of bDMARDs initiation was compared across the clusters over the follow-up period (Fig. 3). In the first 30 months, the probability for bDMARD initiation was significantly different between the subgroups (log-rank $p<0.05$) and the hazard ratio of the C1 subgroup was significantly higher than those of the C2, C3 and C4 subgroups (Fig. 4A). However, this difference dissipated over 30 months (log-rank $p>0.05$), and the hazard ratio of the C1 subgroup plateaued while those of the other subgroups showed a gradual increase (Fig. 4A).

To identify significant predictors of bDMARD initiation in the first 30 months, clinical variables were entered to the multivariable Cox proportional hazards regression model (Fig. 4B). Overweight, smoking, ESR, RF, ACPA, and DAS28 were significantly associated with initiation of bDMARDs. A positive association of smoking and DAS28 (HR [95% CI] 1.51 [1.09–2.08], $p=0.012$ and HR [95% CI] 1.29 [1.09–1.54], $p=0.004$) was in accordance with a previous study (24). A high titre of ACPA was positively associated with the initiation of bDMARDs (HR [95% CI] 1.67 [1.14–1.54], $p=0.008$), while high-titre RF had a negative association (HR [95% CI] 0.66 [0.46–0.95], $p=0.027$).

For significant predictors of bDMARD initiation, the proportion of categorical variables and the mean value (± standard errors) of the continuous variables were compared across the clusters (Fig. 5). In the C1 subgroup, the proportion of smokers, RF (high titre) and ACPA (high titre) were largest and DAS28 was highest. In addition, this subgroup had the second largest rate of being overweight and the second highest ESR

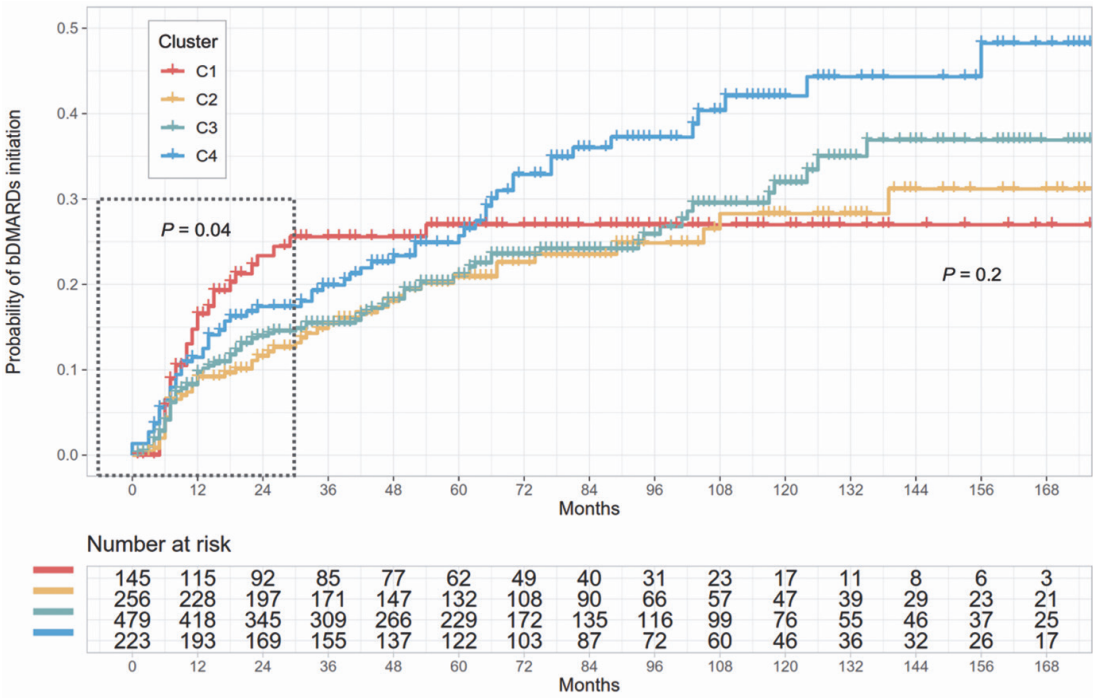

Fig. 3. Cumulative incidence of bDMARD initiation according to cluster. The difference for bDMARD initiation between clusters was significant ($p<0.05$ by the log-rank test) in the first 30 months (dotted-line box).

level. These results suggest that the C1 subgroup has a combination of high-risk factors for bDMARD initiation.

Discussion

In the current study, we demonstrated the feasibility of a novel subclassification of patients with RA. These results were obtained from 1,103 patients with early RA. We analysed the clinical and phenotypic data of RA patients in an unbiased manner using unsupervised learning. We successfully divided them into four mutually exclusive subgroups in terms of clinical features and recapitulated this with a clinically applicable decision tree tool. The identified subgroups have different risk profiles and probability for initiation of bDMARDs, a clinical index of poor long-term outcome.

We entered 14 clinical parameters that characterise RA or have prognostic values into a *k*-means clustering algorithm and obtained four optimally segregated clusters of patients with RA who had distinct clinical features. Subclassification of RA was feasible by a four-step division using four parameters (sex, smoking, hypertension, and dyslipidaemia). Sex and smoking are well-known factors associated with disease activity, treatment response, and clinical outcome in RA (25-27). Hypertension can be linked to inflammation and autoimmunity directly or via salt intake (28). Hypertension is recognised to be a state of chronic inflammation with elevated levels of inflammatory cytokines and with activation of the immune system (28). An increase in blood pressure leads to mechanical and oxidative damage in the endothelial cells, resulting in the formation of danger-associated molecular patterns (DAMPs) such as high mobility group box 1, mitochondrial DNA and heat shock proteins 70 (29). DAMPs are detected by Toll-like receptors and NOD-like receptors in macrophages and fibroblast-like synoviocytes, which produce pro-inflammatory cytokines (IL-1b and IL-18) and chemokines (CCL2 and CCL5) in response. This could boost the synovial inflammation (30). Salt may directly influence the release of inflammatory cytokines such as TNF-α and IL-6 (31, 32). Recent studies reported that increased salt concentration induces serum glucocorticoid kinase 1 (SGK1) expression, which enhances production of interleukin-17-producing CD4+ helper T cells (T_H17) (33, 34). These T cells are highly proinflammatory (33, 34) and may contribute to the pathogenesis of RA (35, 36). In fact, high sodium intake was associated with the risk of RA and ACPA positivity, particularly in smokers (37, 38). Cellular and molecular links between autoimmunity and lipid metabolism were also documented (39). Innate immune cells, including macrophages and dendritic cells, sense lipid species, such as saturated fatty acids and oxidised low-density lipoprotein (LDL), and produce pro-inflammatory cytokines and chemokines. In particular, autoreactive Th17 cell differentiation is augmented under pro-atherogenic condition or stimulation of oxidized LDL, indicating that dysregulated lipid metabolism could affect the pathogenesis of autoimmune disease via Th17 cell (40). It is known that the plasticity of Th17 cells in RA is closely associated with the pathogenicity and disease activity of RA (41). In this regard, complex and close interaction between the four clinical variables and disease-specific features of RA was considered to be merged into the distinct four clusters, and this may facilitate the stratification of RA patients using these four variables.

This study showed that patients in the C1 subgroup were more likely to initiate bDMARDs in the early phase of RA than those in the C2, C3, and C4 subgroups. The C1 subgroup was characterised as male, current smokers, and with elevated CRP and high-titre autoantibodies. Patients in the C1 subgroup shared poor prognostic factors shown in previous studies. In the European League Against Rheumatism (EULAR)

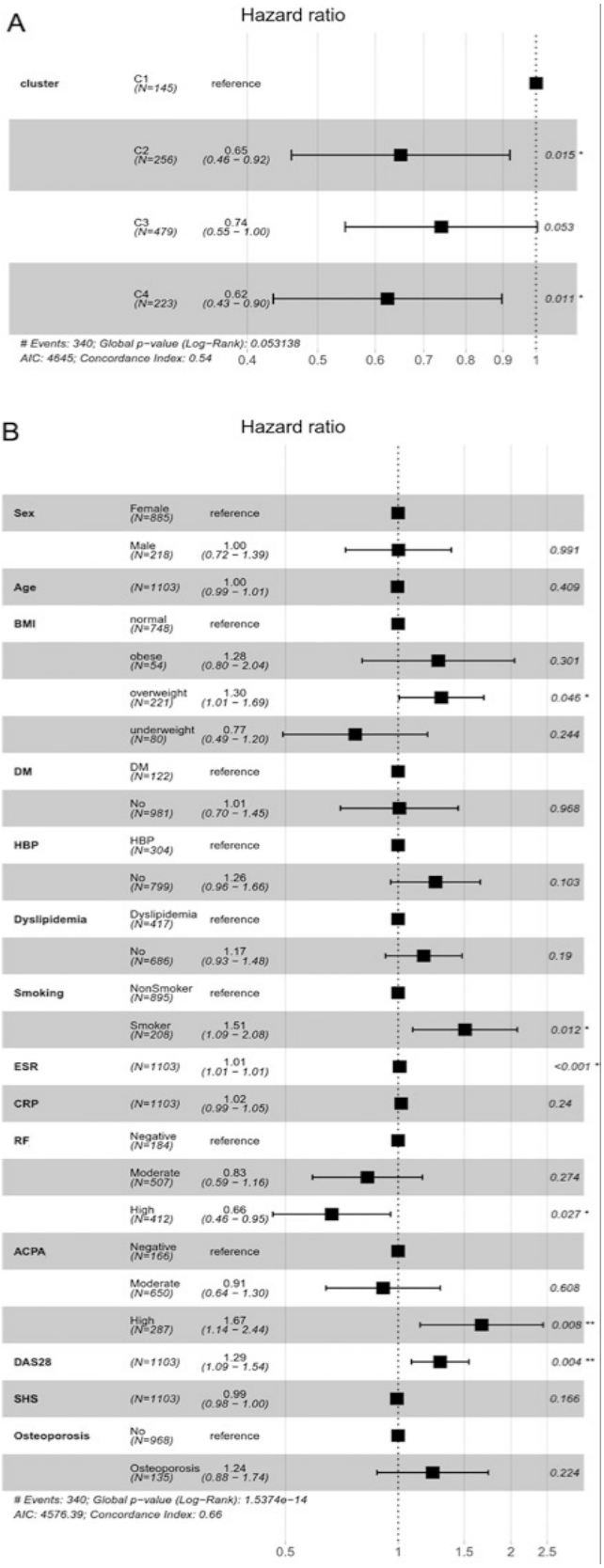

Fig. 4. Forest plots showing hazard ratios of bDMARD initiation in patients with RA. **A**: Hazard ratio by assigned clusters. **B**: Hazard ratio by associated clinical variables.
The Cox proportional hazards model was used to estimate hazard ratios and associated 95% confidence intervals.

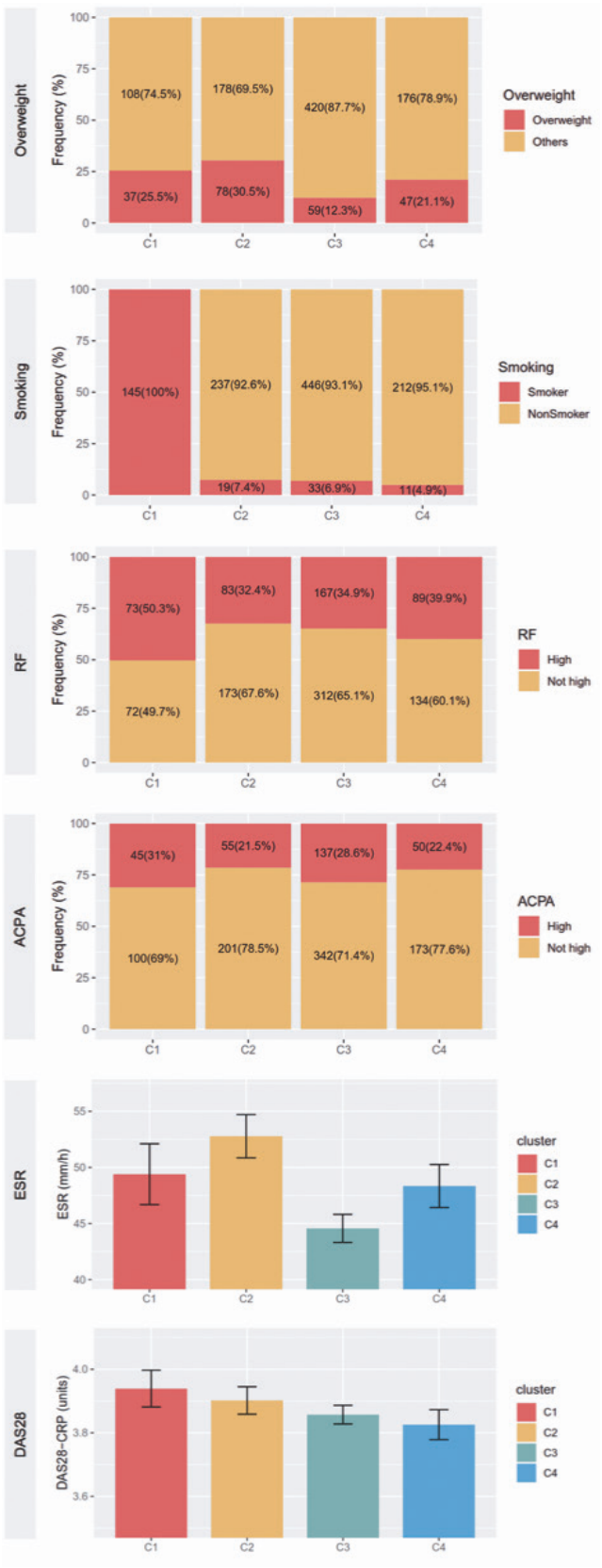

Fig. 5. Comparison of predictors for bDMARD initiation according to the cluster. For the six significant predictors of bDMARD initiation in multivariable Cox regression analysis, the categorical and continuous variables are presented as percentage and mean values (± standard errors), respectively.

recommendations, the poor prognostic factors of RA are defined as moderate to high disease activity despite cDMARD therapy, high levels of ESR or CRP, high swollen joint counts and the presence of high titres of autoantibodies (42). If the therapeutic target is not achieved after initial cDMARD therapy, patients with these poor prognostic factors are recommended to start bDMARDs rather than switch to different cDMARDs. According to the definition of the EULAR recommendations, the C1 subgroup has combinations of poor prognostic factors, and may have a high risk of bDMARD initiation (42). In addition, the current study showed that overweight, smoking, ESR, DAS28, and autoantibodies were independent predictors of bDMARD initiation in the first 30 months. As described previously, the prognostic significance of disease activity, smoking, and autoantibodies on clinical outcome of RA have been supported by previous studies. Being overweight or obese was also associated with a lower chance of achieving the therapeutic target in patients with RA (43, 44). Overweight or obese patients with RA showed a higher degree of synovitis with an abundance of CD68+ CD20+ inflammatory cells, in the early phase of RA as well as during DAS28-based clinical remission (45). The analysis of each subgroup revealed that the C1 subgroup had a high-risk for bDMARDs initiation. This is in good line with the EULAR recommendation for the management of RA (42).

It was reported that RF is a predictor of radiographic progression and poor outcome (46, 47). However, in our results, RF of a high titre was inversely associated with the initiation of bDMARDs. Modern treatment strategies may undermine the impact of RF status as a prognostic marker (48). We also cannot exclude the possibility of index event bias (49, 50). When multiple risk factors contribute to the risk of an outcome, conditioning of the outcome induces dependence between the risk factors, even when these risk factors are independently distributed in the general population. This effect can weaken the existing association or create a spurious association among these risk factors with an index event. It is considered that more powerful predictors such as ACPA, DAS28, and smoking might overwhelm the RF in multivariable analysis. Nevertheless, the C1 high-risk subgroup had the highest proportion of patients with high-titre RF. Despite the growing knowledge of prognostic indicators in RA, it is not easy to calculate the weight of various clinical variables and subsequently evaluate the risk in individual patients. This study demonstrates the possibility of using a simple classification to predict the prognosis of RA. This classification is useful because RA patients are classified into four distinct subgroups using only four factors that are readily identifiable in clinical practice.

Although there are uniform criteria for the classification of RA, the clinical presentation and treatment outcome can be heterogeneous. Clustering of RA patients through machine learning enables a systematic approach stratifying patients with various clinical profiles. In inflammatory diseases, a disease subgroup with a different clinical course often has different molecular or pathophysiologic backgrounds (5, 6, 8, 10). Correlation between synovial tissue signature and treatment response in RA patients suggests that different molecular mechanisms underlie different clinical phenotypes (51-53). Thus, this phenotype-based clustering may provide a good starting point for gaining insight into the divergent mechanistic features of RA. Recently, there have been requirements to implement data science in the field of rheumatic disease. Soon, this computational approach may enable the prediction of the therapeutic response and apply individualised treatments to patients with RA (54).

There are several limitations in this study. First, the clinical data was collected retrospectively. Due to the inherent limitation of retrospective data collection, there may be a selection bias or attrition bias. Second, patients were not treated equally at RA onset. As treatment of RA was determined independently based on the shared decision between patient and physician, selection of DMARDs could be affected by the patient's and physician's preference, as well as economic considerations. However, all patients were treated by rheumatologists, and bDMARDs were initiated based on the contemporary clinical guidelines. Third, the prevalence of smoking, hypertension, and dyslipidaemia varies depending on the ethnicity and region. As these variables are determinants in the subclassification of RA, the distribution of subclass may differ in other cohorts. However, machine learning provides unbiased clustering of patients with distinct phenotypes, and the prognostic utility of subclassification was identified by analysing further clinical progression. It is expected that integrative analysis combining clinical, genetic, molecular, pathologic, and environmental profiles should provide the better answer to this challenge in future.

In the current study, we subclassified patients with RA into four distinct clusters using four clinical variables identified through machine learning. The C1 subgroup characterised by combinations of poor prognostic factors was associated with an early initiation of bDMARDs. This subclassification of heterogeneous patients with RA may provide the opportunities to predict the clinical course and to provide personalised treatments to individual patients.

Acknowledgement

We appreciate the Editage from Cactus for the English editing service.

References

1. SMOLEN JS, ALETAHA D, BARTON A *et al*.: Rheumatoid arthritis. *Nat Rev Dis Primers* 2018; 4: 18001.
2. DEANE KD, DEMORUELLE MK, KELMENSON LB, KUHN KA, NORRIS JM, HOLERS VM: Genetic and environmental risk factors for rheumatoid arthritis. *Best Pract Res Clin Rheumatol* 2017; 31: 3-18.
3. TOWNSEND MJ: Molecular and cellular heterogeneity in the Rheumatoid Arthritis synovium: clinical correlates of synovitis. *Best Pract Res Clin Rheumatol* 2014; 28: 539-49.
4. CONIGLIARO P, TRIGGIANESE P, DE MARTINO E *et al*.: Challenges in the treatment of Rheumatoid Arthritis. *Autoimmun Rev* 2019; 18: 706-13.
5. MOORE WC, MEYERS DA, WENZEL SE *et al*.: Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 2010; 181: 315-23.
6. HOWRYLAK JA, FUHLBRIGGE AL, STRUNK RC, ZEIGER RS, WEISS ST, RABY BA: Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. *J Allergy Clin Immunol* 2014; 133: 1289-300, 300.e1-12.

7. AHMAD T, LUND LH, RAO P *et al*.: Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *J Am Heart Assoc* 2018; 7.
8. GANDELMAN JS, BYRNE MT, MISTRY AM *et al*.: Machine learning reveals chronic graft-versus-host disease phenotypes and stratifies survival after stem cell transplant for hematologic malignancies. *Haematologica* 2019; 104: 189-96.
9. SEYMOUR CW, KENNEDY JN, WANG S *et al*.: Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* 2019; 321: 2003-17.
10. WEISER M, SIMON JM, KOCHAR B *et al*.: Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut* 2018; 67: 36-42.
11. NEOGI T, ALETAHA D, SILMAN AJ *et al*.: The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: Phase 2 methodological report. *Arthritis Rheum* 2010; 62: 2582-91.
12. VAN RIEL PL, FRANSEN J: DAS28: a useful instrument to monitor infliximab treatment in patients with rheumatoid arthritis. *Arthritis Care Res* (Hoboken) 2005; 7: 189-90.
13. SYVERSEN SW, GAARDER PI, GOLL GL *et al*.: High anti-cyclic citrullinated peptide levels and an algorithm of four variables predict radiographic progression in patients with rheumatoid arthritis: results from a 10-year longitudinal study. *Ann Rheum Dis* 2008; 67: 212-7.
14. HECHT C, ENGLBRECHT M, RECH J *et al*.: Additive effect of anti-citrullinated protein antibodies and rheumatoid factor on bone erosions in patients with RA. *Ann Rheum Dis* 2015; 74: 2151-6.
15. VAN DER HEIJDE D: How to read radiographs according to the Sharp/van der Heijde method. *J Rheumatol* 2000; 27: 261-3.
16. JAIN AK, MURTY MN, FLYNN PJ: Data clustering: a review. *ACM Comput Surv* 1999; 31: 264-323.
17. MEHAR AM, MATAWIE K, MAEDER A: Determining an optimal value of K in K-means clustering. 2013 IEEE International Conference on Bioinformatics and Biomedicine. 2013: 51-5.
18. ALELYANI S, TANG J, LIU H: Feature selection for clustering: A review. Data Clustering. Chapman and Hall/CRC, 2018: 29-60.
19. ALI BB, MASSMOUDI Y: K-means clustering based on gower similarity coefficient: A comparative study. 2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO).2013: 1-5.
20. MAATEN LVD, HINTON G: Visualizing data using t-SNE. *J Machine Learning Res* 2008; 9: 2579-605.
21. LOH W-Y: Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2011; 1: 14-23.
22. THERNEAU T, ATKINSON B: rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15 ed. 2019.
23. LEWIS RJ: An introduction to classification and regression tree (CART) analysis. Annual meeting of the society for academic emergency medicine in San Francisco, California.2000.
24. TEITSMA XM, JACOBS JWG, WELSING PMJ *et al*.: Inadequate response to treat-to-target methotrexate therapy in patients with new-onset rheumatoid arthritis: development and validation of clinical predictors.
25. SOKKA T, TOLOZA S, CUTOLO M *et al*.: Women, men, and rheumatoid arthritis: analyses of disease activity, disease characteristics, and treatments in the QUEST-RA study. *Arthritis Res Ther* 2009; 11: R7.
26. MA MH, IBRAHIM F, WALKER D *et al*.: Remission in early rheumatoid arthritis: predicting treatment response. *J Rheumatol* 2012; 39: 470-5.
27. VITTECOQ O, RICHARD L, BANSE C, LE-QUERRE T: The impact of smoking on rheumatoid arthritis outcomes. *Joint Bone Spine* 2018; 85: 135-8.
28. AFSAR B, KUWABARA M, ORTIZ A *et al*.: Salt intake and immunity. *Hypertension* 2018; 72: 19-23.
29. DRUMMOND GR, VINH A, GUZIK TJ, SOBEY CG: Immune mechanisms of hypertension. *Nat Rev Immunol* 2019; 19: 517-32.
30. ELSHABRAWY HA, ESSANI AE, SZEKANECZ Z, FOX DA, SHAHRARA S: TLRs, future potential therapeutic targets for RA. *Autoimmun Rev* 2017; 16: 103-13.
31. KOSTYK AG, DAHL KM, WYNES MW *et al*.: Regulation of chemokine expression by NaCl occurs independently of cystic fibrosis transmembrane conductance regulator in macrophages. *Am J Pathol* 2006; 169: 12-20.
32. HASHMAT S, RUDEMILLER N, LUND H, ABAIS-BATTAD JM, VAN WHY S, MATTSON DL: Interleukin-6 inhibition attenuates hypertension and associated renal damage in Dahl salt-sensitive rats. *Am J Physiol Renal Physiol* 2016; 311: F555-61.
33. KLEINEWIETFELD M, MANZEL A, TITZE J *et al*.: Sodium chloride drives autoimmune disease by the induction of pathogenic TH17 cells. *Nature* 2013; 496: 518-22.
34. WU C, YOSEF N, THALHAMER T *et al*.: Induction of pathogenic TH17 cells by inducible salt-sensing kinase SGK1. *Nature* 2013; 496: 513-7.
35. JUNG SM, KIM Y, KIM J *et al*.: Sodium chloride aggravates arthritis via Th17 polarization. *Yonsei Med J* 2019; 60: 88-97.
36. VAN HAMBURG JP, TAS SW: Molecular mechanisms underpinning T helper 17 cell heterogeneity and functions in rheumatoid arthritis. *J Autoimmun* 2018; 87: 69-81.
37. SUNDSTROM B, JOHANSSON I, RANTAPAA-DAHLQVIST S: Interaction between dietary sodium and smoking increases the risk for rheumatoid arthritis: results from a nested case-control study. *Rheumatology* (Oxford) 2015; 54: 487-93.
38. JIANG X, SUNDSTROM B, ALFREDSSON L, KLARESKOG L, RANTAPAA-DAHLQVIST S, BENGTSSON C: High sodium chloride consumption enhances the effects of smoking but does not interact with SGK1 polymorphisms in the development of ACPA-positive status in patients with RA. *Ann Rheum Dis* 2016; 75: 943-6.
39. RYU H, LIM H, CHOI G *et al*.: Atherogenic dyslipidemia promotes autoimmune follicular helper T cell responses via IL-27. *Nat Immunol* 2018; 19: 583-93.
40. LIM H, KIM YU, SUN H *et al*.: Proatherogenic conditions promote autoimmune T helper 17 cell responses *in vivo*. *Immunity* 2014; 40: 153-65.
41. YANG P, QIAN FY, ZHANG MF *et al*.: Th17 cell pathogenicity and plasticity in rheumatoid arthritis. *J Leukoc Biol* 2019; 106: 1233-40.
42. SMOLEN JS, LANDEWÉ RB, BIJLSMA JW *et al*.: EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. *Ann Rheum Dis* 2020; 79: 685-99.
43. SANDBERG ME, BENGTSSON C, KALLBERG H *et al*.: Overweight decreases the chance of achieving good response and low disease activity in early rheumatoid arthritis. *Ann Rheum Dis* 2014; 73: 2029-33.
44. SCHULMAN E, BARTLETT SJ, SCHIEIR O *et al*.: Overweight, obesity, and the likelihood of achieving sustained remission in early rheumatoid arthritis: results from a multicenter prospective cohort study. *Arthritis Care Res* (Hoboken) 2018; 70: 1185-91.
45. ALIVERNINI S, TOLUSSO B, GIGANTE MR *et al*.: Overweight/obesity affects histological features and inflammatory gene signature of synovial membrane of rheumatoid arthritis. *Sci Rep* 2019; 9: 10420.
46. ALETAHA D, ALASTI F, SMOLEN JS: Rheumatoid factor determines structural progression of rheumatoid arthritis dependent and independent of disease activity. *Ann Rheum Dis* 2013; 72: 875-80.
47. EDWARDS CJ, KIELY P, ARTHANARI S *et al*.: Predicting disease progression and poor outcomes in patients with moderately active rheumatoid arthritis: a systematic review. *Rheumatol Adv Pract* 2019; 3: rkz002.
48. CARPENTER L, NORTON S, NIKIPHOROU E *et al*.: Reductions in radiographic progression in early rheumatoid arthritis over twenty-five years: changing contribution from rheumatoid factor in two multicenter uk inception cohorts. *Arthritis Care Res* (Hoboken) 2017; 69: 1809-17.
49. CHOI HK, NGUYEN US, NIU J, DANAEI G, ZHANG Y: Selection bias in rheumatic disease research. *Nat Rev Rheumatol* 2014; 10: 403-12.
50. DAHABREH IJ, KENT DM: Index event bias as an explanation for the paradoxes of recurrence risk research. *JAMA* 2011; 305: 822-3.
51. DENNIS G, JR., HOLWEG CT, KUMMERFELD SK *et al*.: Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. *Arthritis Res Ther* 2014; 16: R90.
52. LEWIS MJ, BARNES MR, BLIGHE K *et al*.: Molecular portraits of early rheumatoid arthritis identify clinical and treatment response phenotypes. *Cell Rep* 2019; 28: 2455-70.e5.
53. HUMBY F, LEWIS M, RAMAMOORTHI N *et al*.: Synovial cellular and molecular signatures stratify clinical response to csDMARD therapy and predict radiographic progression in early rheumatoid arthritis patients. *Ann Rheum Dis* 2019; 78: 761-72.
54. SILVAGNI E, GIOLLO A, SAKELLARIOU G *et al*.: One year in review 2020: novelties in the treatment of rheumatoid arthritis. *Clin Exp Rheumatol* 2020; 38: 181-94.