

Novel classification of axial spondyloarthritis to predict radiographic progression using machine learning

Y.B. Joo¹, I.-W. Baek², K.-S. Park¹, I. Tagkopoulos³, K.-J. Kim¹

¹Division of Rheumatology, Department of Internal Medicine, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea;

²Division of Rheumatology, Department of Internal Medicine, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea;

³Department of Computer Science & Genome Center, University of California, Davis, USA.

Abstract Objective

Prediction and determination of drug efficacy for radiographic progression is limited by the heterogeneity inherent in axial spondyloarthritis (axSpA). We investigated whether unbiased clustering analysis of phenotypic data can lead to coherent subgroups of axSpA patients with a distinct risk of radiographic progression.

Methods

A group of 412 patients with axSpA was clustered in an unbiased way using a agglomerative hierarchical clustering method, based on their phenotype mapping. We used a generalised linear model, naïve Bayes, Decision Trees, K-Nearest-Neighbors, and Support Vector Machines to construct a consensus classification method. Radiographic progression over 2 years was assessed using the modified Stoke Ankylosing Spondylitis Spine Score (mSASSS).

Results

axSpA patients were classified into three distinct subgroups with distinct clinical characteristics. Sex, smoking, HLA-B27, baseline mSASSS, uveitis, and peripheral arthritis were the key features that were found to stratifying the phenogroups. The three phenogroups showed distinct differences in radiographic progression rate ($p<0.05$) and the proportion of progressors ($p<0.001$). Phenogroup 2, consisting of male smokers, had the worst radiographic progression, while phenogroup 3, exclusively suffering from uveitis, showed the least radiographic progression. The axSpA phenogroup classification, including its ability to stratify risk, was successfully replicated in an independent validation group.

Conclusion

Phenotype mapping results in a clinically relevant classification of axSpA that is applicable for risk stratification. Novel coupling between phenotypic features and radiographic progression can provide a glimpse into the mechanisms underlying divergent and shared features of axSpA.

Key words

axial spondyloarthritis, radiographic progression, machine learning

Young Bin Joo, MD, PhD

In-Woon Baek, MD

Kyung-Su Park, MD, PhD

Ilias Tagkopoulos, 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,

16247 Gyeonggi-do, Republic of Korea.

E-mail: md21c@catholic.ac.kr

Received on January 10, 2020; accepted in revised form on May 11, 2020.

© Copyright CLINICAL AND EXPERIMENTAL RHEUMATOLOGY 2020.

Introduction

Axial spondyloarthritis (axSpA) is a chronic, progressive disease characterised by inflammation of entheses, leading to new bone formation and ankylosis of joints, primarily in the axial skeleton (1, 2). axSpA exhibits a distinct combination of clinical manifestations: sacroiliitis, facet joint arthritis, syndesmophyte formation, and bamboo-like ankylosis of the spine, and often involves peripheral arthritis or enthesitis (1, 2). Interestingly, axSpA is intimately linked to extraarticular manifestations such as uveitis, psoriasis, and inflammatory bowel disease (1, 2). However, in contrast to well-defined phenotypic characteristics, no definite serological biomarkers for diagnosis and subclassification have been discovered except for HLA-B27, which is a strong genetic risk factor.

Most of the rheumatic diseases are not defined as a single entity but classified as a single group according to the defined classification criteria (3, 4). Accordingly, efforts have been made to further classify the disease, such as rheumatoid arthritis, systemic lupus erythematosus, and systemic sclerosis, based on histologic features, molecular signatures, and serological markers to better understand these diseases and tailor intervention or therapy to individuals or subgroups (5-10). However, only limited attempts have been made to further classify axSpA (11, 12), e.g. into HLA-B27 positive versus negative and familial versus sporadic groups. This is ascribed partially to the unavailability of the target tissue (inflammatory enthesis of spine or facet joints) from the patients, no practical biomarker, and the inherent heterogeneity of axSpA, which has a wide variety of clinical manifestations and progression.

Recent advances in machine learning have successfully tackled multi-dimensional heterogeneous data, of similar nature to those collected in the case of axSpA patients (13-16). Prior other studies in diseases such as heart failure and graft-vs-host disease have successfully coupled phenotypic characteristics and their clinical outcome with machine learning approaches (15, 16). In a similar fashion, we hypothesise that applying

machine learning algorithms to clinical phenotyping will enable the detection of novel patterns in multidimensional data obtained from patients with axSpA. We further hypothesise that the identified phenogroups of patients with axSpA will have distinct clinical profiles and differential radiographic progression outcomes. We therefore investigated the utility of unbiased phenotype mapping algorithms in patients with axSpA and validated its functional ability to predict radiographic progression in an independent group of axSpA patients.

Methods

Patients

A total of 412 axSpA patients who fulfilled the Assessment of Spondyloarthritis International Society (ASAS) classification criteria for axSpA (17, 18), and had received care at St. Vincent's Hospital, The Catholic University of Korea (Suwon, Republic of Korea) between 2008 and 2017 were identified. The patients who were clinically suspected of having sacroiliitis but had no definite evidence of it on radiographs or magnetic resonance images were excluded according to the ASAS classification criteria. Clinical, laboratory data and radiographic images were retrieved from the medical records. At baseline, sex, age at diagnosis, time since diagnosis, HLA-B27 status, smoking status, and history of extraarticular manifestations (uveitis, psoriasis, inflammatory bowel disease, peripheral arthritis and enthesitis) were recorded. Disease activity was assessed according to the Ankylosing Spondylitis Disease Activity Score (ASDAS) using C-reactive protein (CRP) level (19). Of these, the 253 patients who were followed up over 2 years were assessed for radiographic progression. For validation analysis, the data of an independent group of 173 patients with axSpA were imported from the Yeouido St. Mary's Hospital, The Catholic University of Korea (Seoul, Republic of Korea) between 2008 and 2017. These patients fulfilled the ASAS classification criteria for axSpA and their clinical profiles were compiled entirely independently. The study was carried out in accordance with the Helsinki Declara-

Competing interests: none declared.

tion and approved by the Institutional Review Board of St. Vincent's Hospital, The Catholic University of Korea (no. VC18RESI0248).

Radiographs and scoring

Radiographs of the sacroiliac joints and the cervical and lumbar spine were obtained by the local investigator at baseline and after 2 years of follow-up. All available radiographs per patient were independently scored at the same time according to the modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) (20) by two experienced readers independently, blinded to all other data except radiograph chronology. Spinal radiographic progression was defined as worsening of the mean mSASSS by more than 2 units over 2 years, in conformity with previous studies (21-23). Radiographic sacroiliitis (SI) was scored according to the modified New York criteria (24) and radiological hip involvement was graded based on the BASRI-hip scoring system (25). The interobserver reliability was assessed by calculating the interclass correlation coefficient (ICC). Agreement between the 2 readers regarding mSASSS status was fair (ICC 0.946 [95% CI 0.940 to 0.952]) and agreement regarding change in the mSASSS was good (ICC 0.692 [95% CI 0.654 to 0.726]). The ICC for sacroiliac and hip joint status grade was 0.889 [95% CI 0.875 to 0.901] and 0.958 [95% CI 0.953 to 0.963]), respectively.

Unsupervised clustering and determination of the optimal number of clusters

Unsupervised learning identifies hidden structure or patterns purely from the information in the data without the need for a training set, classes or labels. To classify the axSpA patients into subgroups based on their phenotypic variables, we used the agglomerative hierarchical clustering method, a commonly used unsupervised learning tool (26). An agglomerative approach begins with each observation in a distinct cluster. Then, the similarity (or distance) between each of the clusters is computed and the two most similar clusters are merged into one. It success-

sively repeats to merges clusters together and updates the proximity matrix until only a single cluster remains. To ensure equal weights of all features in their presentation, the 23 representative phenotypic variables were standardised within a range of 0 to 1. Agglomerative hierarchical clustering was performed with the dissimilarity matrix given by Euclidean distance (27) and the average linkage score was used to join similar clusters. The Euclidean distance is the ordinary straight-line distance between two points in Euclidean space, and the larger the distance between two clusters, the more distinct it is. The average linkage criterion involves looking at the distances between all pairs and averages all of these distances. To determine the optimal number of phenogroups within the axSpA group, we used model-based clustering, which assumes a Gaussian distribution for values of phenotypic variables within a cluster and achieves parameter fitting and patient assignment by maximising the penalised likelihood (28). To do this, we used the mclust package in R and explored a full range of covariance structures, some of which relax the requirement for feature independence (*i.e.* non-diagonal covariance matrixes are allowed). The Bayesian information criterion (BIC) was used to penalise increases in model complexity such as a greater number of clusters or variability in the standard deviation across variables and across clusters (29). The BIC procedure is to choose the model for which the BIC criterion is maximised, and in our implementation, we investigated models of one to five clusters.

To confirm unsupervised clustering results, we used *t*-distributed stochastic neighborhood embedding (*t*-SNE) (30), a powerful dimensionality reduction method. 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.

Ensemble of supervised learning for the novel phenogroups

Given a set of data, a supervised learning algorithm attempts to optimise a

function (the model) to find the combination of feature values that result in the target output. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model to decrease variance, bias, or improve predictions. To obtain consensus-based classification for the phenogroups, we constructed an ensemble of five classifiers: a generalised linear model (GLM), naïve Bayes (NB), decision trees (DT), K-nearest-neighbours (KNN), and support vector machines (SVM) (31, 32). In consensus-based classification, the final phenogroup label is the one with the highest number of votes. For KNN, the *k* parameter was tuned in the range 3 to 20. The SVM was built using a Gaussian radial basis function kernel and the sigma hyperparameter was determined from an estimation based upon the 0.1 and 0.9 quantiles of the samples. For the soft margins, the *C* parameter that achieved the best performance was in the range of 2^{-4} to 2^7 . To avoid the overfitting and minimise the error, we applied a 10-fold cross-validation for each model. Model fitting was performed using the caret package in R.

Selection of informative features

To capture the features that provide the salient information that distinguishes phenogroups, we employed the wrapper feature selection method (32, 33). The wrapper method uses the classifier as a black box to rank different subsets of the features according to their predictive power.

Statistical analyses

For continuous distributed data, the results are shown as means with standard deviation (SD); between-group comparisons were performed using Student's *t*-test or ANOVA. Categorical or dichotomous variables are expressed as frequencies and percentages, and were compared using the chi-squared test or Fisher's exact test. Multivariable logistic regression analysis was performed to identify independent predictors associated with radiographic progression. Survival rates and corresponding 95% confidence intervals were estimated by Kaplan-Meier

Table I. Baseline characteristics of the study subjects and comparison between phenogroups.

Variable	All patients (n=412)	Phenogroup 1 (n=182)	Phenogroup 2 (n=158)	Phenogroup 3 (n=72)	p-value
Male, n (%)	302 (73.3)	96 (52.7)	158 (100.0)	48 (66.7)	<0.001
Age, years	36.5 ± 12.5	36.9 ± 12.8	35.1 ± 12.6	38.6 ± 11.4	0.134
Age at diagnosis, years	33.2 ± 12.0	33.4 ± 12.3	32.1 ± 11.9	34.9 ± 11.7	0.243
Disease duration, years	4.0 ± 5.2	3.61 ± 4.95	4.31 ± 5.51	4.38 ± 5.19	0.204
BMI, kg/m ²	23.8 ± 3.7	22.7 ± 3.4	23.9 ± 3.3	23.6 ± 3.9	0.005
HLA-B27, n (%)	344 (83.5)	123 (67.6)	156 (98.7)	70 (97.2)	<0.001
Smoking					<0.001
Never, n (%)	211 (54.2)	133 (73.1)	52 (32.9)	35 (48.6)	
Ex-smoker, n (%)	40 (10.3)	14 (7.7)	22 (13.9)	5 (6.9)	
Current smoker, n (%)	138 (35.5)	35 (19.2)	84 (53.2)	32 (44.4)	
Peripheral arthritis, n (%)	131 (31.8)	118 (64.8)	3 (1.9)	10 (13.9)	<0.001
Enthesitis, n (%)	24 (5.8)	20 (11.0)	4 (2.5)	0 (0.0)	<0.001
Uveitis, n (%)	76 (18.4)	4 (2.2)	2 (1.3)	70 (97.2)	<0.001
Psoriasis, n (%)	10 (2.4)	8 (4.4)	1 (0.6)	1 (1.4)	<0.001
Inflammatory bowel disease, n (%)	9 (2.2)	2 (1.1)	6 (3.8)	1 (1.4)	0.208
ESR, mm/hr	36.5 ± 27.8	42.4 ± 31.0	30.7 ± 23.9	33.5 ± 23.7	<0.001
CRP, mg/dL	1.9 ± 3.0	2.3 ± 3.6	1.4 ± 1.8	1.8 ± 3.4	0.035
ASDAS-CRP	2.6 ± 0.9	2.6 ± 1.0	2.5 ± 0.8	2.5 ± 0.9	0.221
Use of TNF inhibitor, n (%) *	202 (49.0)	98 (53.8)	70 (44.4)	34 (47.2)	0.202
Sacroiliitis according to mNY criteria	222 (53.8)	86 (47.3)	91 (57.6)	45 (62.5)	0.044
mSASSS, units	8.4 ± 14.6	6.9 ± 14.2	10.5 ± 15.3	8.1 ± 13.8	0.070
Cervical spine	3.9 ± 7.4	3.3 ± 7.3	4.0 ± 7.2	3.6 ± 6.7	0.687
Lumbar spine	4.9 ± 9.0	3.6 ± 8.0	6.5 ± 9.9	4.5 ± 8.4	0.008
Presence of syndesmophyte(s), n (%)	173 (42.0)	125 (68.7)	92 (58.2)	43 (59.7)	0.107
Number of syndesmophyte(s)	2.6 ± 5.0	2.2 ± 5.0	3.2 ± 5.2	2.4 ± 4.7	0.147
Cervical spine	1.2 ± 2.6	1.1 ± 2.6	1.3 ± 2.6	1.1 ± 2.4	0.719
Lumbar spine	1.4 ± 3.0	1.0 ± 2.7	1.9 ± 3.4	1.3 ± 2.8	0.030
Sacroiliac joint, n (%)					0.067 / 0.037
Grade 0 (Right / Left)	77 (18.7) / 87 (21.1)	41 (22.5) / 51 (28.0)	25 (15.8) / 23 (14.6)	11 (15.3) / 13 (18.1)	
Grade 1 (Right / Left)	93 (22.6) / 85 (20.6)	41 (22.5) / 37 (20.3)	38 (24.1) / 36 (22.8)	14 (19.4) / 12 (16.7)	
Grade 2 (Right / Left)	112 (27.2) / 97 (23.5)	56 (30.8) / 43 (23.6)	38 (24.1) / 34 (21.5)	18 (25.0) / 20 (27.8)	
Grade 3 (Right / Left)	57 (13.8) / 77 (18.7)	25 (13.7) / 32 (17.6)	21 (13.4) / 33 (20.9)	11 (15.3) / 12 (16.7)	
Grade 4 (Right / Left)	73 (17.7) / 66 (16.0)	19 (10.4) / 19 (10.4)	36 (22.8) / 32 (20.3)	18 (25.0) / 15 (20.8)	
Hip involvement, n (%)					0.545 / 0.292
Grade 0 (Right / Left)	368 (89.3) / 368 (89.3)	159 (87.4) / 162 (89.0)	145 (91.8) / 141 (89.2)	64 (88.9) / 65 (90.3)	
Grade 1 (Right / Left)	25 (6.1) / 28 (6.8)	15 (8.2) / 15 (8.2)	5 (3.2) / 8 (5.1)	5 (6.9) / 5 (6.9)	
Grade 2 (Right / Left)	14 (3.4) / 9 (2.2)	7 (3.8) / 2 (1.1)	5 (3.2) / 5 (3.2)	2 (2.8) / 2 (2.8)	
Grade 3 (Right / Left)	2 (0.5) / 4 (0.9)	1 (0.5) / 3 (1.6)	1 (0.6) / 1 (0.0)	0 (0.0) / 0 (0.0)	
Grade 4 (Right / Left)	3 (0.7) / 3 (0.7)	0 (0.0) / 0 (0.0)	2 (1.3) / 3 (1.9)	1 (1.4) / 0 (0.0)	
BMD, g/cm ²					
Lumbar spine	1.060 ± 0.152	1.060 ± 0.161	1.060 ± 0.153	1.040 ± 0.122	0.709
Femoral neck	0.852 ± 0.118	0.840 ± 0.102	0.864 ± 0.144	0.854 ± 0.084	0.168
Total hip	0.910 ± 0.116	0.891 ± 0.103	0.928 ± 0.133	0.920 ± 0.100	0.011
Z score					
Lumbar spine	-0.243 ± 1.140	-0.187 ± 1.220	-0.289 ± 1.150	-0.286 ± 0.903	0.670
Femoral neck	-0.399 ± 0.848	-0.455 ± 0.788	-0.401 ± 0.929	-0.253 ± 0.801	0.231
Total hip	-0.442 ± 0.813	-0.521 ± 0.717	-0.397 ± 0.938	-0.340 ± 0.734	0.187

ASDAS: Ankylosing Spondylitis Disease Activity Score; BMD: bone mineral density; BMI: body mass index; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; Lt: left; mNY: modified New York; Rt: right; TNF: tumour necrosis factor.

* Counted if ever used. TNF inhibitors include etanercept, adalimumab, infliximab, and golimumab.

analysis and compared using log-rank tests. The risk of discontinuation of TNF inhibitor treatment was evaluated using a 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. 3.5.1, The R Project for Statistical Computing, www.r-project.org).

Results

Clustering of the study population and their characteristics

The baseline characteristics of the study subjects (n=412) are given in Table I and an overview of the analysis steps is given in Fig. 1. Because there is redundancy of information among some variables, we selected the following 23 representative features for clustering

analysis: sex, age at diagnosis, disease duration, body mass index (BMI), HLA-B27 status, smoking status, peripheral arthritis, enthesitis, uveitis, psoriasis, inflammatory bowel disease, erythrocyte sedimentation rate (ESR), CRP, ASDAS-CRP, modified Stoke Ankylosing Spondylitis Spine Score (mSASSS), syndesmophyte(s), sacroiliac joint grades (right and left), hip joint grades

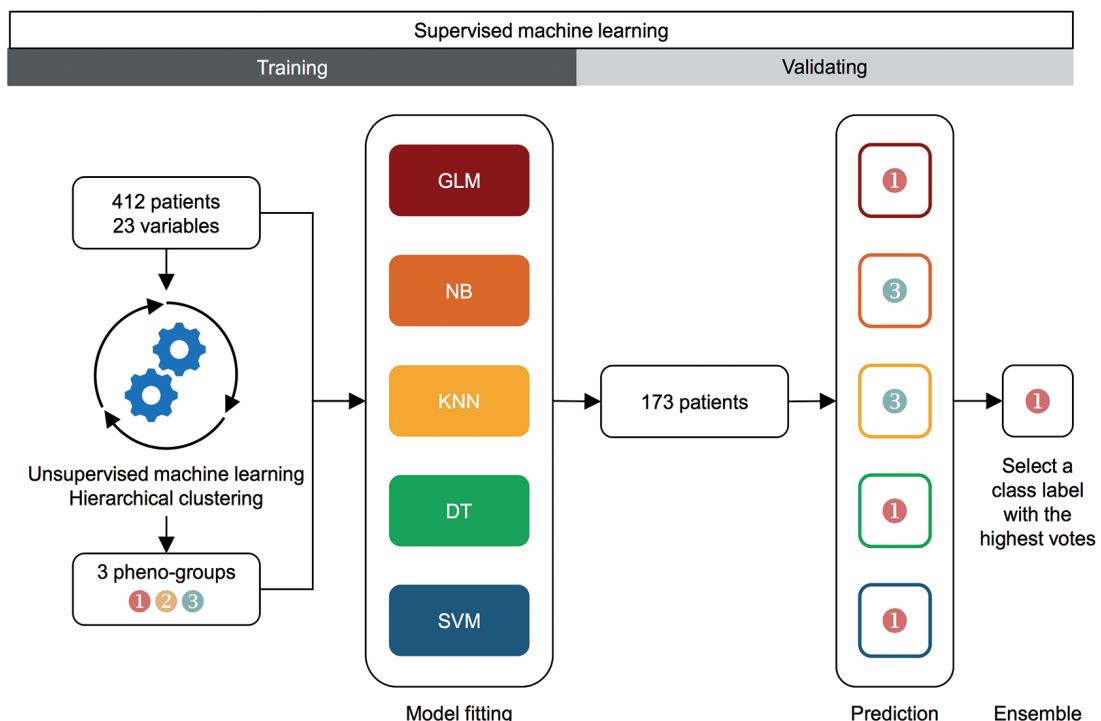


Fig. 1. Overview of the data analysis steps. A training dataset of 412 patients and 23 clinical variables was clustered in an unbiased way using agglomerative hierarchical clustering, which is an unsupervised machine learning method, resulting in three phenogroups with distinct features. Phenotypic variables and the assigned labels (phenogroups) were learned using an ensemble model consisting of five classifiers (GLM, NB, KNN, DT, and SVM). A validation dataset of 173 patients and 23 variables was predicted using the trained models and the class label with the highest number of votes is given to each observation. DT, decision tree; GLM, generalised linear model; KNN, k-nearest neighbours; NB, naïve Bayesian; SVM, support vector machine.

(right and left), and Z-score (lumbar spine, femur neck, and total hip). Use of the TNF inhibitor was excluded because it is not a phenotype but a treatment method to control disease activity.

The axSpA patients were then categorised into subgroups based on their phenotypic profiles using model-based hierarchical clustering (28). To identify the optimal number of clusters and to assess robustness of the clustering result, we computed the Bayesian information criterion for different numbers of clusters from 1 to 5, where we found that 3 clusters are the optimal representation of the data (Fig. 2A) and dendrogram of 3 clusters is shown in Fig. 2B. Phenogroup was assigned in their order of membership and the phenotype heat map given in Fig. 2C. Segregation of axSpA phenogroups was also reproduced by *t*-distributed stochastic neighbourhood embedding (*t*-SNE), which is an unsupervised machine learning algorithm that project all patients onto a two-dimensional plane by reducing dimensionality (Fig. 2D) (30).

The characteristics of the phenogroups are compared in Table I. Phenogroup

1 is composed of approximately equal numbers of men and women who were mostly non-smokers (73.1%) and the least positive for HLA-B27 (67.6%). In contrast, phenogroup 2 is all male and has the highest rate of ex- or current-smokers (67.1%). Phenogroup 3 is a mixture of men and women in a ratio of 2:1 and the ex- or current-smoker rate is 51.3%. No significant difference exists in baseline age, age at diagnosis, and disease duration across phenogroups. Peripheral arthritis and enthesitis are much more common in phenogroup 1, while uveitis is exclusively predominant in phenogroup 3. Psoriasis is also most common in phenogroup 1 although low in frequency (4.4%). Mean ESR and CRP is highest in phenogroup 1, but ASDAS does not differ across phenogroups ($p=0.221$). Radiographic indices of spine (mSASSS and number of syndesmophyte) is marginally higher in phenogroup 2 than phenogroups 1 and 3 ($p=0.070$ and $p=0.147$, respectively), and the proportion of severe SI and hip injury is highest in phenogroup 2. The BMD and Z-score of the lumbar spine and femur neck are comparable across

the phenogroups. The use of the TNF inhibitor does not differ between the phenogroups ($p=0.202$). These results demonstrate that axSpA can be further classified into subgroups with distinct features.

Association of phenogroups with radiographic progression

For 253 of the 412 patients, radiographs in the spine were available at baseline and at 2-year follow-up. There are no significant differences in the baseline characteristics between a subset of the 2-year follow-up and original populations (Supplementary Table S1). We examined the radiographic progression over 2 years according to phenogroup. The rate of increase in mSASSS is significantly higher in phenogroup 2 than phenogroup 1 and 3 (1.19 ± 1.94 vs. 0.664 ± 1.27 and 0.548 ± 1.37 , all $p<0.05$; Fig. 3A). For the change in the number of syndesmophytes, phenogroup 2 shows a higher increase than phenogroups 2 and 3 (0.905 ± 1.73 vs. 0.482 ± 1.16 and 0.432 ± 0.219 , all $p<0.05$; Fig. 3B). The proportion of progressors in phenogroup 2 is higher

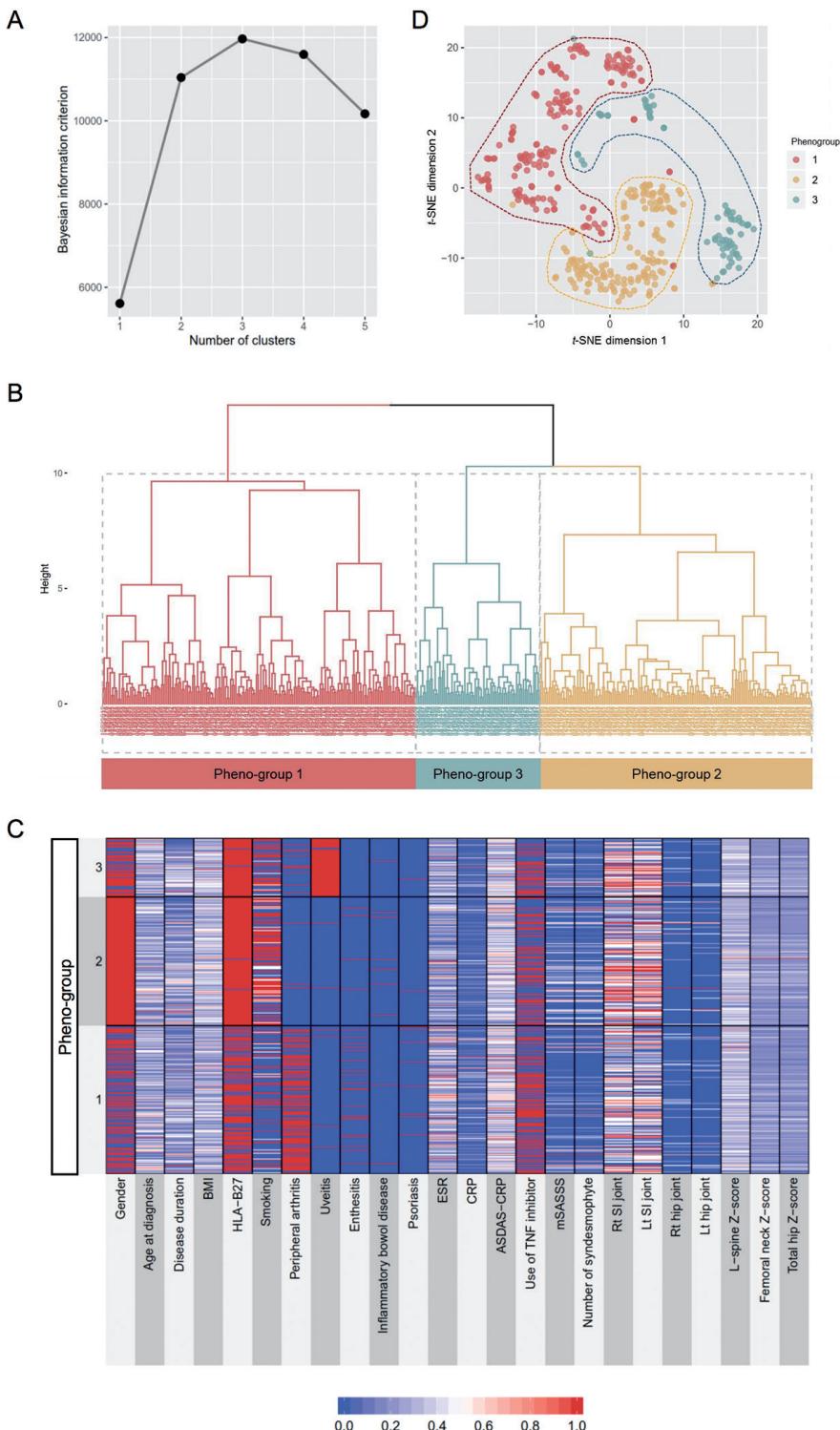


Fig. 2. A: Bayesian information criterion analysis for the identification of the optimal number of phenogroups. B: Dendrogram of the hierarchical clustering based on the phenotypic features. C: Heatmap of the hierarchical clustering. Columns are the clinical and phenotypic variables and rows represent the clustered individual study participants. Red indicates a higher value of a phenotype whereas blue indicates a lower value. Dichotomous or graded variables are designated as discrete scales. For example, sex, HLA-B27, or uveitis are expressed as 1 (*i.e.* male, positive, or presence) or 0 (*i.e.* female, negative, or absence). (D) t-SNE plot of the phenogroups.

than that of phenogroup 3 (33.7% vs. 11.4%, $p<0.05$), but insignificant with respect to phenogroup 1 (33.7% vs.

23.7%, $p=0.148$; Fig. 3C). However, the proportion of progressors definitely showed a decreasing tendency for phe-

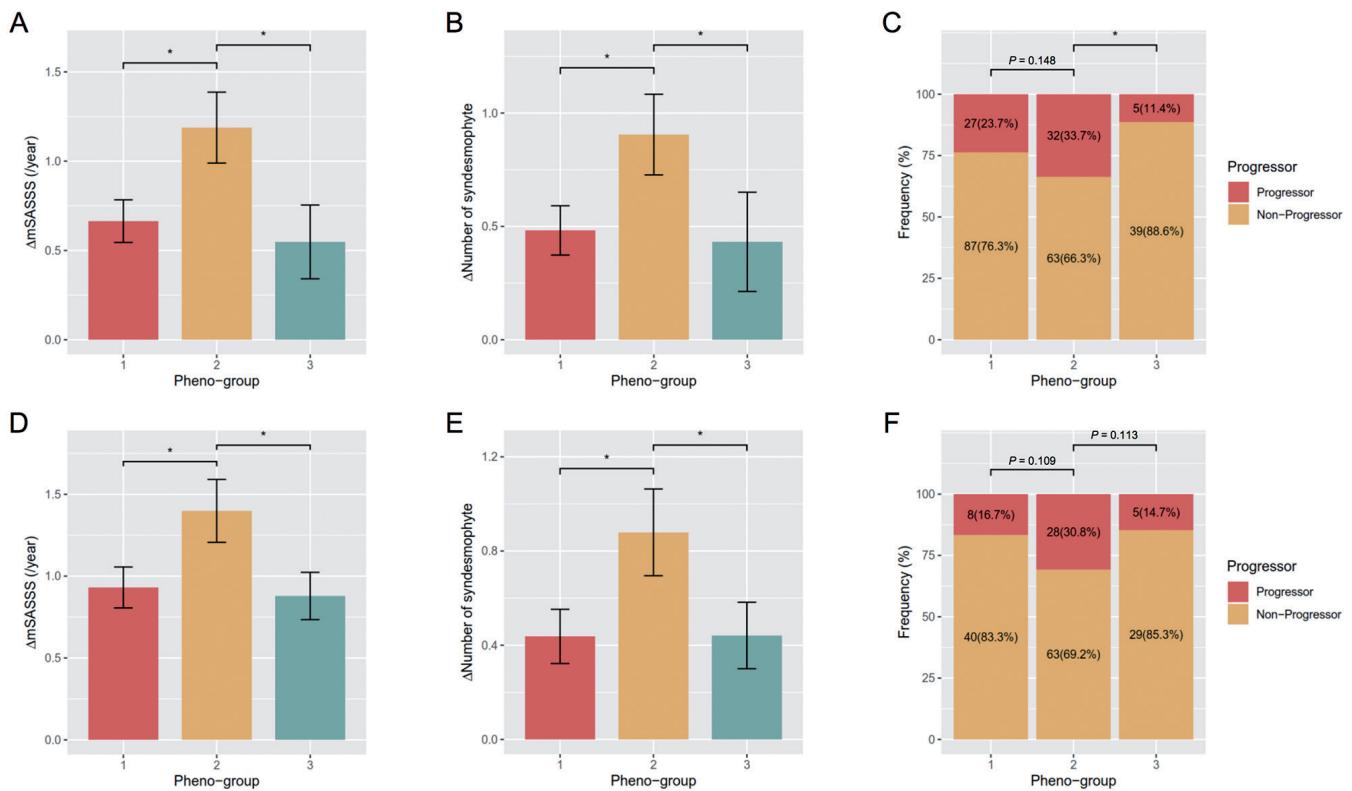
nogroups 2, 1, and 3 in that order (p for trend <0.001).

To confirm the association between the phenogroups and radiographic progression, we performed multivariable logistic regression analysis including the phenogroups in addition to known predictors (Table II). Phenogroups 1 and 2 are significantly associated with radiographic progression in comparison with phenogroup 3 (phenogroup 1: OR 6.36, 95% CI, 1.61 to 25.05, $p=0.008$; phenogroup 2: OR 6.36, 95% CI, 1.82 to 22.14, $p=0.004$). The presence of syndesmophyte(s) at baseline, male sex, body mass index, and baseline ESR are also independently associated with radiographic progression, and this is consistent with previous results (21, 23, 34, 35).

Validation of the phenogroups analyses

To validate our results, we imported the data of an independent 173 patients with axSpA from an affiliated hospital. The characteristics of the validation group are compared in Supplementary Table S2. In the validation group, the proportion of male patients is higher (83.8% vs. 73.3%, $p=0.009$) and peripheral arthritis is less frequent (15.6% vs. 31.8%, $p<0.001$). The distribution of hip joint grade is marginally different (right and left, $p=0.053$ and 0.048, respectively), but there is no difference in the level of severity (grades 3 and 4, $p=0.905$). The higher BMD and Z scores in the validation group seem to be ascribable to the higher male proportion. No significant difference between the two groups was found in the other variables.

We constructed a consensus classification scheme that consists of five different classifiers: a generalised linear model (GLM), naïve Bayes (NB), decision trees (DT), K-nearest-neighbours (KNN), and support vector machines (SVM) (Fig. 1). Each model was fitted to the phenotypic features and the assigned classes (phenogroups) of the training dataset. The key features that discriminate the phenogroups were explored using the wrapper method for each model. Sex, smoking, uveitis, peripheral arthritis, and HLA-B27 are the

**Fig. 3.** Radiographic progression over 2 years by phenogroup.

A: Change in mSASSS in the training dataset. **B:** Change in number of syndesmophytes in the training dataset. **C:** Proportion of progressors in the training dataset. **D:** Change in mSASSS in the validation dataset. **E:** Change in number of syndesmophytes in the validation dataset. **F:** Proportion of progressors in the validation dataset. * indicates $p < 0.05$.

Table II. Association between the novel phenogroups and radiographic progression in patients with axial spondyloarthritis.

Variables	Crude OR (95% CI)	Adjusted OR (95% CI)	p-value
Phenogroup (ref. phenogroup 3)			
1	2.42 (0.87, 6.76)	6.36 (1.61, 25.05)	0.008
2	3.96 (1.42, 11.03)	6.34 (1.82, 22.14)	0.004
Male sex	5.62 (2.15, 14.72)	4.87 (1.35, 17.53)	0.010
Age at diagnosis	1.04 (1.01, 1.06)	1.01 (0.97, 1.05)	0.568
Disease duration	1.03 (0.99, 1.07)	1.03 (0.97, 1.10)	0.317
BMI	1.88 (1.25, 2.84)	2.31 (1.32, 4.03)	0.003
HLA-B27	1.02 (0.45, 2.30)	0.84 (0.24, 2.98)	0.791
Smoking	2.27 (1.27, 4.03)	1.83 (0.81, 4.13)	0.143
ESR	1.01 (1.1, 1.02)	1.02 (1.1, 1.04)	0.017
CRP	1.07 (0.98, 1.18)	1.02 (0.86, 1.20)	0.861
ASDAS-CRP	1.35 (0.99, 1.84)	0.83 (0.47, 1.49)	0.536
Presence of syndesmophyte(s) at baseline	10.91 (5.47, 21.73)	8.42 (3.32, 21.35)	<0.001
Use of TNF inhibitor	0.7 (0.4, 1.24)	0.59 (0.27, 1.27)	0.175

ASDAS: Ankylosing Spondylitis Disease Activity Score; BMI: body mass index; CI: confidence interval; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; OR: odds ratio; TNF: tumour necrosis factor.

key features of the KNN, DT, and SVM models. Uveitis, peripheral arthritis, HLA-B27, and mSASSS are the informative variables of the GLM model. In contrast, uveitis, peripheral arthritis, and sex are the informative variables of the NB model.

The fitted models were applied to the validation group to predict phenogroup. The phenogroup with the highest number of votes of the five models was assigned to each patient in the validation group. Phenogroup 2 was most often predicted ($n=91$, 52.6%), followed by

phenogroup 1 ($n=48$, 27.7%) and phenogroup 3 ($n=34$, 19.7%). Each of the predicted phenogroups was successfully matched to the pre-defined phenogroups (Supplementary Table S3). We examined the radiographic progression over 2 years with respect to the predicted phenogroups in the validation group. The rate of increase in mSASSS is significantly higher in phenogroup 2 than phenogroup 1 and 3 (1.40 ± 1.83 vs. 0.931 ± 0.866 and 0.879 ± 0.843 , all $p < 0.05$; Fig. 3D). Phenogroup 2 is also superior to phenogroup 2 and 3 with respect to the change in number of syndesmophytes (0.879 ± 1.76 vs 0.438 ± 0.796 and 0.441 ± 0.824 , all $p < 0.05$; Fig. 3E). The proportion of progressors is marginally higher in phenogroup 2 than in phenogroups 1 and 3 (30.8% vs. 16.7% and 14.7%, $p=0.109$ and 0.113, respectively; Fig. 3F). However, the proportion of progressors clearly shows a decreasing tendency phenogroups 2, 1, and 3 in that order (p for trend =0.030). This result confirms that the phenogroups have a different risk for radiographic progression.

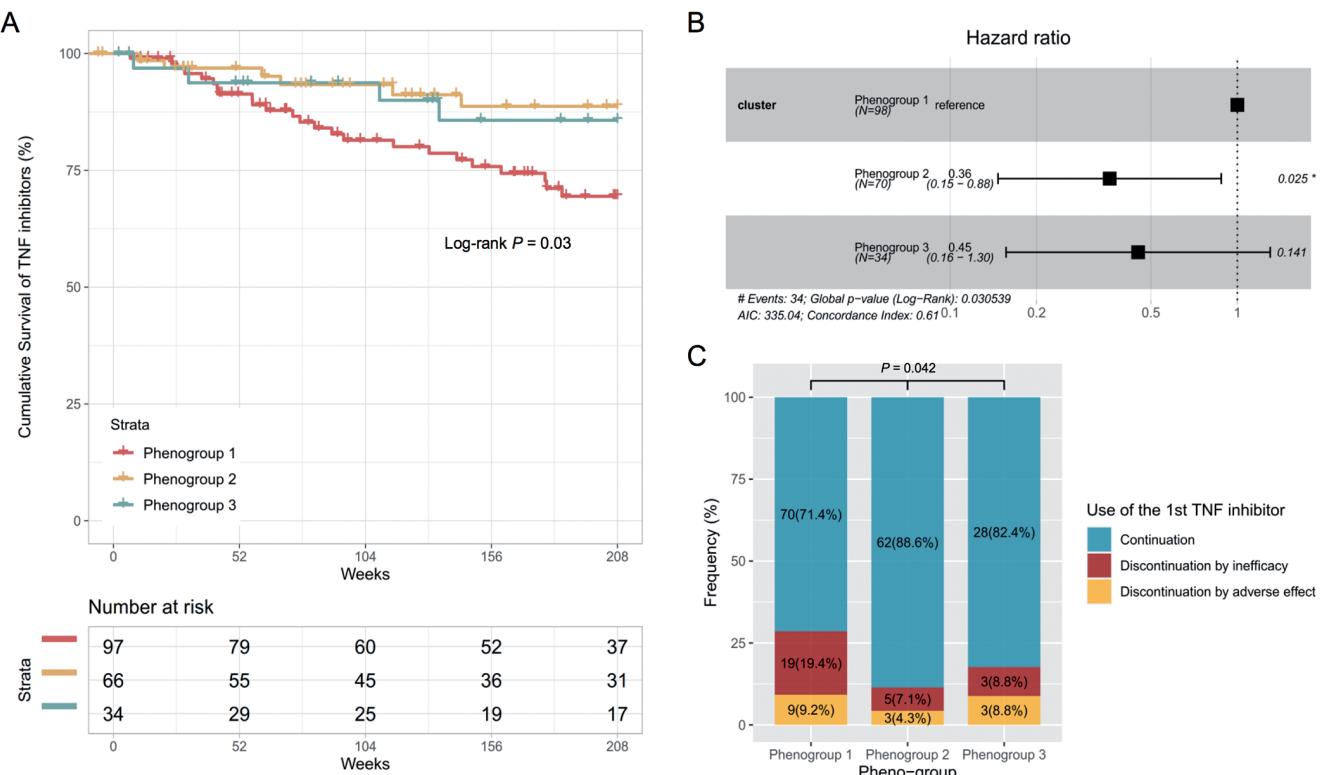


Fig. 4. Drug survival rate and risk of discontinuation of the first TNF inhibitor over 4 years by phenogroup.

A: Cumulative survival curves of the first TNF inhibitor derived by Kaplan-Meier analysis. **B:** Forest plot showing hazard ratios of TNF inhibitor discontinuation created using a Cox proportional hazards regression model. **C:** Frequency of TNF inhibitor use and discontinuation, including reasons for discontinuation. Statistical analysis using a chi-square test was conducted to evaluate the differences in the proportions of each phenogroup.

Risk of discontinuation of the first TNF inhibitors

Drug survival is a composite measure of effectiveness and safety (36, 37). Thus, we further investigated the survival rate of the first TNF inhibitor in the three subgroups over 4 years (208 weeks). Results showed that the drug retention rate differed between the three phenogroups (Log-rank $p=0.03$): it was high in phenogroups 2 and 3 but low in phenogroup 1 (Fig. 4A). Compared with phenogroup 1, the risk of discontinuing TNF inhibitor treatment was substantially lower in phenogroup 2 (HR [95% CI] 0.36 [0.15–0.88], $p=0.025$) and in phenogroup 3 (HR [95% CI] 0.45 [0.16–1.30], $p=0.141$) (Fig. 4B). Reasons for discontinuing TNF inhibitor treatment included inefficacy (including primary and secondary inefficacy) (64.3% of cases, $n=27$) and adverse effects (35.7% of cases, $n=15$) (Fig. 4C). The rate of discontinuation owing to inefficacy was highest in phenogroup 1 (19.4%, $p=0.004$), while the rate of continuation was highest in phenogroup 2 (88.6%).

Discussion

In the current study, we demonstrated the feasibility and validity of a novel classification technique for axSpA, a heterogeneous group of rheumatic diseases. These results were obtained for a group of 412 patients with axSpA along with a validation group of 173 independent patients with axSpA. We analysed the phenotypic data of axSpA patients in an unbiased manner using unsupervised learning and successfully divided them into three mutually exclusive phenogroups in term of clinical features. The identified phenogroups have differential radiographic progression, indicating differing risk profiles and clinical trajectories.

AxSpA is a well-characterised disease with a combination of unique spinal and extraspinal features, and the scoring systems measuring the disease activity and structural change of the involved sites are established. These systems enabled the phenotype mapping and a novel grouping. Although all patients met the established classification criteria for axSpA, the mapped phenotypes

clearly demonstrate that axSpA is heterogeneous in nature.

The three identified phenogroups showed striking differences. Phenogroup 1 is a mixed group with members that are more associated with peripheral arthritis, enthesitis, and psoriasis. Members in phenogroup 1 are mostly non-smokers. Phenogroup 2 is a subset of males with few extraspinal manifestations such as peripheral arthritis, enthesitis, and uveitis. Phenogroup 2 has the highest ex- or current-smoker rate (67.1%). Phenogroup 3 is a subpopulation of mixed males and females in a ratio of 2:1 with an exclusively high rate of uveitis. The advanced stages of SI and hip joint are more common in phenogroup 2. Phenogroup 2 had a higher mSASSS at baseline than phenogroup 1 and 3, and this is probably due to the high proportion of males and smokers, which are known unfavourable risk factors for radiographic progression (1, 38). The key features stratifying the phenogroups by machine learning models are sex, smoking, HLA-B27, baseline mSASSS, uveitis, and peripheral

arthritis. It is known that uveitis is more prevalent in HLA-B27-positive patients than HLA-B27-negative patients (39). In our result, HLA-B27-positive patients are further divided into two subgroups (phenogroups 2 and 3) by the presence of uveitis. A close link between females and non-smokers in phenogroup 1 is ascribed in part to the low smoking rate (5% to 7%) of females in Korea (40). The low rate of psoriasis (7.4%) in Asian axSpA patients (37) could also alter the effect of psoriasis on the grouping.

As an independent measure of how well these groups are separated, we clinically validated the association of phenogroups with radiographic progression, demonstrating the robust potential of phenogroup membership as a method for risk stratification. Factors negatively affecting the radiographic progression in axSpA are well known (1, 38). Inarguably, the patients in phenogroup 2, with its unfavourable risk factors of male sex, smoking, and syndesmophytes at baseline, showed the worst radiographic progression rate of the phenogroups. In contrast, the patients in phenogroup 1, with its favourable risk factors of female sex, no smoking, and fewer syndesmophytes at baseline, exhibited a retarded radiographic progression. An interesting subset is phenogroup 3, which showed the least radiographic progression although its patients had moderate risk profiles compared with those of phenogroups 1 and 2. Clinical features and risk factors for radiographic progression in axSpA are well-defined but it remains undefined how many subgroups would be distinguishable by the reasonable method. This result demonstrated that a group of axSpA patients with heterogeneity and high dimensional variables was successfully reduced into three subgroups by calculating the similarities and distance between patients. Unravelling the intriguing coupling between phenogroup and radiographic progression could help us better understand the aetiopathogenesis of axSpA.

The protective effect of TNF inhibitors on radiographic progression or new bone formation in axSpA remains controversial. Three open-label exten-

sions of randomised controlled trials of TNF inhibitors in axSpA over 2 years failed to demonstrate an inhibition of radiographic progression in comparison with a historical cohort of patients not treated with TNF inhibitors (41–43). In contrast, use of a TNF inhibitor was associated with a reduction of radiographic progression in a 2-year prospective observational study (44). These conflicting results could have several reasons and one could be driven by stratification given the heterogeneity of a study population that has different risk profiles. In the latter study, when the patients were categorised into three subgroups according to ASDAS level, it was found that radiographic progression was significantly impeded for those patients with a lower ASDAS level at baseline (44). In this context, our study has several important ramifications for the design of future axSpA clinical trials and the research of axSpA. Future clinical trials can take into account the phenotypic categorisation by performing the deep phenotyping of study participants and differential analysis for radiographic progression. This approach can be also used in a clinical setting to determine whether certain groups of patients are more responsive to the investigational drug than other types of patients. Although epidemiological studies and observational registries of axSpA have enrolled a wide variety of patients with varying phenotypic characters, detailed mechanistic studies of axSpA are often selective on the patient phenotype, hence limiting their ability to generalise their findings to the broader axSpA patient population. In some inflammatory diseases, a disease subgroup with a different clinical outcome has different clinical phenotypes or molecular backgrounds (13, 14, 16, 45). This phenotype-based clustering may provide a good starting point for catching a glimpse into the divergent mechanistic features of axSpA. It would be also interesting to investigate whether the distinct subgroups show a different response to TNF inhibitors or investigational drugs.

It is worthy of note that the rate of discontinuation owing to inefficacy was highest in phenogroup 1, while the rate

of continuation was highest in phenogroup 2. This finding indicates that patients in phenogroup 1 were more resistant to TNF inhibitors for reasons that remain unclear. However, while patients in phenogroup 2 responded relatively well to TNF inhibitors, and the level of the TNF inhibitor was maintained over time, these patients exhibited worse radiographic progression. These findings imply that more well-established, mature inflammatory lesions may be present in the spines of patients in phenogroup 2, as indicated in the TNF-brake hypothesis (46). Thus, we hypothesise that IL-17 inhibitors, which show promising efficacy in suppressing new bone formation, would be more suitable for patients in phenogroup 2 (47, 48). Independent validation steps in large patient cohorts with multimodal approaches are needed to validate this hypothesis (49).

There are some limitations in this study that should be addressed. First, data were retrospectively collected. Retrospective data collection is inherently susceptible to bias, including both misclassification and information bias. Second, a variation in the prevalence of extraarticular manifestations has been found across different races and geographical areas. Because uveitis and peripheral arthritis are the key features stratifying the patients, phenotype mapping and clustering could differ in other cohorts. However, the primary goal of our study is to show the use of an unbiased approach allows the clustering of patients into distinct and mutually exclusive groups that could be used in the clinic or in clinical trials, not to define specific subgroups. Third, the effect of TNF inhibitors on radiographic progression was not fully considered, although the frequency in use of TNF inhibitors did not differ across the 3 phenogroups and treatment durations are less than 2 years. Fourth, functional metrics such as BASFI or BASMI were not included in the phenotype mapping because of a lack of data. However, spinal mobility or functional activity is strongly dependent on inflammation and potentially reversible (50, 51).

In the present study, we demonstrated how one can use unbiased cluster

analysis of phenotypic data to obtain clinically relevant categories of axSpA patients with significantly different risk of radiographic progression. In consideration of the heterogeneity of axSpA, phenotype mapping could be useful for the subclassification of axSpA patients and may enable a more personalised therapy in the future. Moreover, phenotype mapping could give information for the design of future clinical trials to better assess their outcomes.

Acknowledgment

We thank Kimberly Moravec, PhD, from Edanz Group (www.edanzediting.com/ac) for editing this manuscript.

References

- SIEPER J, BRAUN J, DOUGADOS M, BAETEN D: Axial spondyloarthritis. *Nat Rev Dis Primers* 2015; 1: 15013.
- SIEPER J, PODDUBNYY D: Axial spondyloarthritis. *Lancet* 2017; 390: 73-84.
- GOLDBLATT F, O'NEILL SG: Clinical aspects of autoimmune rheumatic diseases. *Lancet* 2013; 382: 797-808.
- SMOLEN JS, ALETAHA D, MCINNES IB: Rheumatoid arthritis. *Lancet* 2016; 388: 2023-38.
- THURLINGS RM, WIJBRANDTS CA, MEBIUS RE et al.: Synovial lymphoid neogenesis does not define a specific clinical rheumatoid arthritis phenotype. *Arthritis Rheum* 2008; 58: 1582-9.
- MILANO A, PENDERGRASS SA, SARGENT JL et al.: Molecular subsets in the gene expression signatures of scleroderma skin. *PLoS One* 2008; 3: e2696.
- BOS CL, VAN BAARSEN LG, TIMMER TC et al.: Molecular subtypes of systemic sclerosis in association with anti-centromere antibodies and digital ulcers. *Genes Immun* 2009; 10: 210-8.
- DE ROOY DP, WILLEMZE A, MERTENS B, HUIZINGA TW, VAN DER HELM-VAN MIL AH: Can anti-cyclic citrullinated peptide antibody-negative RA be subdivided into clinical subphenotypes? *Arthritis Res Ther* 2011; 13: R180.
- ORANGE DE, AGIUS P, DICARLO EF et al.: Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. *Arthritis Rheumatol* 2018; 70: 690-701.
- YU F, HAAS M, GLASSOCK R, ZHAO MH: Redefining lupus nephritis: clinical implications of pathophysiological subtypes. *Nat Rev Nephrol* 2017; 13: 483-95.
- ALMODOVAR R, FONT P, ZARCO-MONTEJO P et al.: Phenotypic differences between familial versus sporadic ankylosing spondylitis: a cross-sectional Spanish registry of spondyloarthropathies (REGISPOSER). *Clin Exp Rheumatol* 2011; 29: 822-7.
- AREVALO M, GRATACOS MASMITJA J, MORENO M et al.: Influence of HLA-B27 on the Ankylosing Spondylitis phenotype: results from the REGISPOSER database. *Arthritis Res Ther* 2018; 20: 221.
- 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.
- 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.
- 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.
- 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.
- RUDWALEIT M, VAN DER HEIJDE D, LANDEWÉ R et al.: The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. *Ann Rheum Dis* 2009; 68: 777-83.
- MEGHNATHI B, SARAU X A, DOUGADOS M, MOLTÓ A: Evaluation of the predictive validity of the ASAS axial spondyloarthritis criteria in the DESIR cohort. *Clin Exp Rheumatol* 2019; 37: 797-802.
- MOLTÓ A, GOSSEC L, MEGHNATHI B et al.: An Assessment in SpondyloArthritis International Society (ASAS)-endorsed definition of clinically important worsening in axial spondyloarthritis based on ASDAS. *Ann Rheum Dis* 2018; 77: 124-7.
- CREEMERS MC, FRANSSEN MJ, VAN'T HOF MA, GRIBNAU FW, VAN DE PUTTE LB, VAN RIEL PL: Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. *Ann Rheum Dis* 2005; 64: 127-9.
- PODDUBNYY D, HAIBEL H, LISTING J et al.: Baseline radiographic damage, elevated acute-phase reactant levels, and cigarette smoking status predict spinal radiographic progression in early axial spondylarthritis. *Arthritis Rheum* 2012; 64: 1388-98.
- RAMIRO S, VAN TUBERGEN A, STOLWIJK C et al.: Scoring radiographic progression in ankylosing spondylitis: should we use the modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) or the Radiographic Ankylosing Spondylitis Spinal Score (RASSS)? *Arthritis Res Ther* 2013; 15: R14.
- PODDUBNYY D, PROTOPOPOV M, HAIBEL H, BRAUN J, RUDWALEIT M, SIEPER J: High disease activity according to the Ankylosing Spondylitis Disease Activity Score is associated with accelerated radiographic spinal progression in patients with early axial spondyloarthritis: results from the GERman SPondyloarthritis Inception Cohort. *Ann Rheum Dis* 2016; 75: 2114-8.
- VAN DER LINDEN S, VALKENBURG HA, CATS A: Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheum* 1984; 27: 361-8.
- MACKAY K, BROPHY S, MACK C, DORAN M, CALIN A: The development and validation of a radiographic grading system for the hip in ankylosing spondylitis: the bath ankylosing spondylitis radiology hip index. *J Rheumatol* 2000; 27: 2866-72.
- MURTAGH F, CONTRERAS P: Methods of hierarchical clustering. *CoRR* 2011; abs/1105.0121.
- MURTAGH F, LEGENDRE PJOC: Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J Classif* 2014; 31: 274-95.
- FRALEY C, RAFTERY AE: Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002; 97: 611-31.
- HASTIE T, TIBSHIRANI R, FRIEDMAN J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer New York, 2013.
- MAATEN LVD, HINTON G: Visualizing data using t-SNE. *J Machine Learning Res* 2008; 9: 2579-605.
- JAMES G, WITTEN D, HASTIE T, TIBSHIRANI R: An Introduction to Statistical Learning: with Applications in R. Springer New York, 2013.
- KUHN M, JOHNSON K: Applied Predictive Modeling. Springer New York, 2013.
- GUYON I, ELISSEFF A: An introduction to variable and feature selection. *J Machine Learning Res* 2003; 3: 1157-82.
- VAN TUBERGEN A, RAMIRO S, VAN DER HEIJDE D, DOUGADOS M, MIELANTS H, LANDEWÉ R: Development of new syndesmophytes and bridges in ankylosing spondylitis and their predictors: a longitudinal study. *Ann Rheum Dis* 2012; 71: 518-23.
- JEONG H, BEA EK, LEE J, KOH EM, CHA HS: Body mass index and estrogen predict radiographic progression in the spine in ankylosing spondylitis. *Joint Bone Spine* 2015; 82: 473-4.
- CARLI L, CALABRESI E, GOVERNATO G, BRAUN J: One year in review 2018: axial spondyloarthritis. *Clin Exp Rheumatol* 2019; 37: 889-98.
- BILGIN E, FARISOĞULLARI B, ARMAĞAN B et al.: Predictors of drug retention and treatment response in axial spondyloarthritis patients treated with certolizumab: real-life results from the HURBIO registry. *Clin Exp Rheumatol* 2019 Aug 14 [Online ahead of print].
- PODDUBNYY D, SIEPER J: Radiographic progression in ankylosing spondylitis/axial spondyloarthritis: how fast and how clinically meaningful? *Curr Opin Rheumatol* 2012; 24: 363-9.
- FELDTKELLER E, KHAN MA, VAN DER HEIJDE D, VAN DER LINDEM S, BRAUN J: Age at disease onset and diagnosis delay in HLAB27 negative vs. positive patients with ankylosing spondylitis. *Rheumatol Int* 2003; 23: 61-6.
- PARK MB, KIM CB, NAM EW, HONG KS: Does South Korea have hidden female smokers: discrepancies in smoking rates between self-reports and urinary cotinine level. *BMC Womens Health* 2014; 14: 156.

41. VAN DER HEIJDE D, LANDEWÉ R, BARALIAKOS X *et al.*: Radiographic findings following two years of infliximab therapy in patients with ankylosing spondylitis. *Arthritis Rheum* 2008; 58: 3063-70.
42. VAN DER HEIJDE D, LANDEWÉ R, EINSTEIN S *et al.*: Radiographic progression of ankylosing spondylitis after up to two years of treatment with etanercept. *Arthritis Rheum* 2008; 58: 1324-31.
43. VAN DER HEIJDE D, SALONEN D, WEISSMAN BN *et al.*: Assessment of radiographic progression in the spines of patients with ankylosing spondylitis treated with adalimumab for up to 2 years. *Arthritis Res Ther* 2009; 11: R127.
44. MOLNAR C, SCHERER A, BARALIAKOS X *et al.*: TNF blockers inhibit spinal radiographic progression in ankylosing spondylitis by reducing disease activity: results from the Swiss Clinical Quality Management cohort. *Ann Rheum Dis* 2018; 77: 63-9.
45. WEISER M, SIMON JM, KOCHAR B *et al.*: Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut* 2018; 67: 36-42.
46. MAKSYMOWYCH WP: Disease modification in ankylosing spondylitis. *Nat Rev Rheumatol* 2010; 6: 75-81.
47. BRAUN J, BARALIAKOS X, DEODHAR A *et al.*: Effect of secukinumab on clinical and radiographic outcomes in ankylosing spondylitis: 2-year results from the randomised phase III MEASURE 1 study. *Ann Rheum Dis* 2017; 76: 1070-7.
48. BRAUN J, HAIBEL H, DE HOOGE M *et al.*: Spinal radiographic progression over 2 years in ankylosing spondylitis patients treated with secukinumab: a historical cohort comparison. *Arthritis Res Ther* 2019; 21: 142.
49. MARCHESONI A, D'ANGELO S, ANZIDEI M *et al.*: Radiologist-rheumatologist multidisciplinary approach in the management of axial spondyloarthritis: a Delphi consensus statement. *Clin Exp Rheumatol* 2019; 37: 575-84.
50. LANDEWÉ R, DOUGADOS M, MIELANTS H, VAN DER TEMPEL H, VAN DER HEIJDE D: Physical function in ankylosing spondylitis is independently determined by both disease activity and radiographic damage of the spine. *Ann Rheum Dis* 2009; 68: 863-7.
51. MACHADO P, LANDEWÉ R, BRAUN J, HERMANN KG, BAKER D, VAN DER HEIJDE D: Both structural damage and inflammation of the spine contribute to impairment of spinal mobility in patients with ankylosing spondylitis. *Ann Rheum Dis* 2010; 69: 1465-70.