Data-driven subclassification of ANCA-associated vasculitis: model-based clustering of a federated international cohort

Karl Gisslander, Arthur White, Louis Aslett, Zdenka Hrušková, Peter Lamprecht, Jacek Musiał, Jamsheela Nazeer, James Ng, Declan O'Sullivan, Xavier Puéchal, Matthew Rutherford, Mårten Segelmark, Benjamin Terrier, Vladimir Tesař, Michelangelo Tesi, Augusto Vaglio, Krzysztof Wójcik, Mark A Little, Aladdin J Mohammad*, on behalf of the FAIRVASC consortium†*

Summary

Background Antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis is a heterogenous autoimmune disease. While traditionally stratified into two conditions, granulomatosis with polyangiitis (GPA) and microscopic polyangiitis (MPA), the subclassification of ANCA-associated vasculitis is subject to continued debate. Here we aim to identify phenotypically distinct subgroups and develop a data-driven subclassification of ANCA-associated vasculitis, using a large real-world dataset.

Methods In the collaborative data reuse project FAIRVASC (Findable, Accessible, Interoperable, Reusable, Vasculitis), registry records of patients with ANCA-associated vasculitis were retrieved from six European vasculitis registries: the Czech Registry of ANCA-associated vasculitis (Czech Republic), the French Vasculitis Study Group Registry (FVSG; France), the Joint Vasculitis Registry in German-speaking Countries (GeVas; Germany), the Polish Vasculitis Registry (POLVAS; Poland), the Irish Rare Kidney Disease Registry (RKD; Ireland), and the Skåne Vasculitis Cohort (Sweden). We performed model-based clustering of 17 mixed-type clinical variables using a parsimonious mixture of two latent Gaussian variable models. Clinical validation of the optimal cluster solution was made through summary statistics of the clusters' demography, phenotypic and serological characteristics, and outcome. The predictive value of models featuring the cluster affiliations were compared with classifications based on clinical diagnosis and ANCA specificity. People with lived experience were involved throughout the FAIRVASVC project.

Findings A total of 3868 patients diagnosed with ANCA-associated vasculitis between Nov 1, 1966, and March 1, 2023, were included in the study across the six registries (Czech Registry n=371, FVSG n=1780, GeVas n=135, POLVAS n=792, RKD n=439, and Skåne Vasculitis Cohort n=351). There were 2434 (62·9%) patients with GPA and 1434 (37·1%) with MPA. Mean age at diagnosis was 57·2 years (SD 16·4); 2006 (51·9%) of 3867 patients were men and 1861 (48·1%) were women. We identified five clusters, with distinct phenotype, biochemical presentation, and disease outcome. Three clusters were characterised by kidney involvement: one severe kidney cluster (555 [14·3%] of 3868 patients) with high C-reactive protein (CRP) and serum creatinine concentrations, and variable ANCA specificity (SK cluster); one myeloperoxidase (MPO)-ANCA-positive kidney involvement cluster (782 [20·2%]) with limited extrarenal disease (MPO-K cluster); and one proteinase 3 (PR3)-ANCA-positive kidney involvement cluster (683 [17·7%]) with widespread extrarenal disease (PR3-K cluster). Two clusters were characterised by relative absence of kidney involvement: one was a predominantly PR3-ANCA-positive cluster (1202 [31·1%]) with inflammatory multisystem disease (IMS cluster), and one was a cluster (646 [16·7%]) with predominantly ear–nose–throat involvement and low CRP, with mainly younger patients (YR cluster). Compared with models fitted with clinical diagnosis or ANCA status, cluster-assigned models demonstrated improved predictive power with respect to both patient and kidney survival.

Interpretation Our study reinforces the view that ANCA-associated vasculitis is not merely a binary construct. Data-driven subclassification of ANCA-associated vasculitis exhibits higher predictive value than current approaches for key outcomes.

Funding European Union's Horizon 2020 research and innovation programme under the European Joint Programme on Rare Diseases.

Introduction

The subclassification of antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis is subject to continued research.[1] The original concept of ANCA-associated vasculitis, combining granulomatosis with polyangiitis (GPA), microscopic polyangiitis (MPA), and eosinophilic granulomatosis with polyangiitis (EGPA), paid tribute to their shared clinical and pathological features and association with positive ANCA serology.[2] However, while classified as ANCA-associated vasculitis, EGPA has less overlap with GPA and MPA than GPA and MPA have with each other with respect to genetics,

Lancet Rheumatol 2024

Published Online August 22, 2024 https://doi.org/10.1016/S2665-9913(24)00187-5

See Online/Comment https://doi.org/10.1016/S2665-9913(24)00264-9

*Contributed equally

†Members are listed in the appendix (p 23)

Rheumatology, Department of Clinical Sciences, Lund University, Lund, Sweden (K Gisslander, A J Mohammad MD PhD); **School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland** (A White PhD, J Ng PhD); **ADAPT Centre, Trinity College Dublin, Dublin, Ireland** (A White, J Nazeer PhD, Prof D O'Sullivan PhD, Prof M A Little MD PhD); **Department of Mathematical Sciences, Durham University, Durham, UK** (L Aslett PhD); **Department of Nephrology, General University Hospital in Prague and First Faculty of Medicine, Charles University, Prague, Czech Republic** (Z Hrušková MD PhD, Prof V Tesař MD PhD); **Department of Rheumatology and Clinical Immunology, University of Lübeck, Lübeck, Germany** (Prof P Lamprecht MD PhD); **II Chair of Internal Medicine, Faculty of Medicine, Jagiellonian University Medical College, Kraków, Poland** (Prof Jacek Musiał MD PhD, K Wójcik MD PhD); **National Referral Center for Rare Systemic Autoimmune Diseases, Hôpital Cochin, Assistance Publique–Hôpitaux de Paris, Université Paris Cité, Paris, France** (Prof X Puéchal MD PhD, B Terrier MD PhD); **French Vasculitis Study Group, Paris, France** (Prof X Puéchal, B Terrier); **School of Infection and Immunity, University of**

Glasgow, Glasgow, UK (M Rutherford MD); **Department of Clinical Sciences, Lund University and Department of Endocrinology, Nephrology and Rheumatology, Skåne University Hospital, Lund, Sweden** (Prof M Segelmark MD PhD); **Nephrology and Dialysis Unit, Azienda Ospedaliera Universitaria Meyer IRCCS, Florence, Italy** (M Tesi, Prof A Vaglio MD PhD); **Department of Biomedical, Experimental and Clinical Sciences "Mario Serio", University of Florence, Florence, Italy** (Prof A Vaglio); **Trinity Kidney Centre, Trinity Translational Medicine Institute, School of Medicine, Trinity College Dublin, Dublin, Ireland** (Prof M A Little); **Department of Medicine, University of Cambridge, Cambridge, UK** (A J Mohammad)

Correspondence to: Karl Gisslander, Rheumatology, Department of Clinical Sciences, Lund University, Lund SE-221 85, Sweden
karl.gisslander@med.lu.se

See Online for appendix

Research in context

Evidence before this study

We searched PubMed using the terms (Anti-Neutrophil Cytoplasmic Antibody-Associated Vasculitis[MeSH Terms]) AND ((Cluster Analysis[MeSH Terms]) OR cluster OR phenotype OR stratification OR subclassification) for studies that aimed to subclassify patients with antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis, published between database inception and May 23, 2024. Two studies of cluster-driven subclassification attempting to cover the full spectrum of disease at the time of diagnosis were identified. One study was done using data from randomised clinical trials. The other study used multicentre, single-nation, observational data. Based on organ system involvement, serology, and laboratory data, the studies generated two widely different phenotypic stratifications. The sample sizes of the two studies were relatively small and the generalisability of the results questionable.

Added value of this study

Our study, based on data from the largest ANCA-associated vasculitis cohort yet reported, reinforces the view that ANCA-associated vasculitis is not merely a binary construct. Using data from six European countries we found five distinct clusters of patients: those with severe kidney disease; those with anti-MPO-positive kidney disease; those with anti-PR3-positive kidney disease; younger patients with low-grade disease activity and respiratory involvement; and those with inflammatory multisystem disease. These clusters had different disease outcomes, and a predictive model including the cluster subclassification was associated with better model performance than existing subclassifications.

Implications of all the available evidence

Our data-driven subclassification of ANCA-associated vasculitis, based on a large real-world dataset, could provide more accurate and clinically useful patient stratification for therapeutic, epidemiological, and basic research than current existing methodologies.

pathophysiology, clinical presentation, and therapy.[3] Consequently, in this Article, further discussion of ANCA-associated vasculitis refers to GPA and MPA, excluding EGPA.

Recently, a reclassification of GPA and MPA has been suggested based on ANCA-specificity, into vasculitis associated with proteinase 3 (PR3)-ANCA and vasculitis associated with myeloperoxidase (MPO)-ANCA. Some evidence suggests that an ANCA-serology-based classification of these subsets better correlates with genetic predisposition,[4] prediction of clinical outcomes,[5] and response to treatment,[6] than clinical diagnosis. However, both dichotomous subclassifications of ANCA-associated vasculitis, either as GPA and MPA, or PR3-ANCA-positive and MPO-ANCA-positive disease, fail to address the considerable heterogeneity seen in disease presentation. To tackle this issue, unbiased data-driven disease subclassification of ANCA-associated vasculitis, without the prespecified notion of distinct disease subtypes or number of groups, has been attempted using unsupervised machine learning.[7–9] However, as ANCA-associated vasculitis is a rare disease, and modern machine learning algorithms are data hungry, small sample sizes have been a limiting factor.

In 2023, the largest cohort of patients with ANCA-associated vasculitis to-date was presented by the FAIRVASC (Findable, Interoperable, Accessible, Reusable, Vasculitis) consortium. In this collaborative project, retrospective observational data from six European registries were quality controlled and harmonised to a common schema.[10]

To further explore the phenotypic spectrum of ANCA-associated vasculitis, we performed an unbiased data-driven subclassification of ANCA-associated vasculitis using model-based clustering, using this large, harmonised cohort of real-world patient data.

Methods

Study design and participants

FAIRVASC is a data reuse project, federating multiple European vasculitis registries through remote data access via a shared interface.[11] In this study we retrieved the harmonised data foundational for FAIRVASC to a central datastore to explore the phenotypic spectrum of ANCA-associated vasculitis. Records of patients with ANCA-associated vasculitis were retrieved from six European vasculitis registries: the Czech Registry of ANCA-associated vasculitis (Czech Republic), the French Vasculitis Study Group Registry (FVSG; France), the Joint Vasculitis Registry in German-speaking Countries (GeVas; Germany), the Polish Vasculitis Registry (POLVAS; Poland), the Irish Rare Kidney Disease Registry (RKD; Ireland) and the Skåne Vasculitis Cohort (Sweden; appendix p 4). These registries recruit patients from secondary or tertiary care or specialised vasculitis centres, or have regional population coverage. The registry structures, data quality, local exclusion criteria, and patient characteristics including periods of recruitment and follow-up of the aggregated patient cohort have previously been described.[10] To enter this study, a new diagnosis of GPA or MPA irrespective of classification criteria applied was required, including a valid date of diagnosis.

This study is part of a data reuse project. Ethical approval for the collection, sharing, and analysis of patient data within the FAIRVASC consortium is

governed locally, and approval was obtained from the ethics committee at each registry site. Written informed consent was obtained directly from patients in five out of six participating registries. In one registry (the Skåne Vasculitis Cohort), informed consent was not obtained and not required according to the ethics committee. The patient perspective has been considered throughout the FAIRVASC project, with three patient representatives involved since inception. Several discussion workshops with different patient groups have been carried out to explore what registries are, what information is stored, data protection and patient rights, and perspectives on research priorities. From a patient perspective, the use of existing information to find patterns of symptoms was considered such a priority.

Procedures

A total of 17 mixed-type variables were used as input for the model. Age at diagnosis, and serum creatinine and C-reactive protein (CRP) concentrations at diagnosis, were regarded as continuous, while gender and the involvement of constitutional symptoms, the musculoskeletal system, skin, eyes, mucosa, ear–nose–throat, lung, the cardiovascular system, the gastrointestinal system, kidneys, and the central and peripheral nervous system at diagnosis were regarded as binary. Gender was assigned as reported in the six source registries; the options available were woman, man, undetermined, unknown, and undifferentiated. ANCA was regarded as a three-level nominal consisting of ANCA negative, PR3-ANCA positive, and MPO-ANCA positive. In the absence of an antigen-specific immunoassay, patients were assigned to PR3-ANCA or MPO-ANCA if immunofluorescence showed cytoplasmic or perinuclear staining pattern, respectively (appendix p 12). Anti-PR3 and anti-MPO double-positive patients were reassigned either PR3-ANCA or MPO-ANCA positivity based on immunofluorescence or clinical diagnosis. Patients undergoing acute dialysis with a missing serum creatinine value were randomly assigned a value ranging from 500 µmol/L to 1000 µmol/L. Due to inconsistencies in variable definitions in the POLVAS registry compared with the other registries, data from this registry on serum creatinine were defined as "highest ever" and data on organ pattern were defined as "at any time". Patients with nine or more missing variables out of the 17 (ie, >50% missing data) were excluded from analysis.

The 17 mixed-type variables and the occurrence of, and time to, end-stage kidney disease and death were used to assess the identified clusters. One registry (POLVAS) did not record the date of end-stage kidney disease and was excluded from this time-to-event analysis. The definition of end-stage kidney disease differed slightly across the registries (appendix p 5).

Statistical analysis

We performed model-based clustering of the mixed data using a parsimonious mixture of two latent Gaussian variable models available in the R package *clustMD*,[12] guided by a framework for multiple imputation in cluster analysis (appendix p 21).[13]

Two mixture models (EVI and VVI) were fitted for G=1–5 clusters using a Monte Carlo expectation maximisation algorithm with 500 iterations. These models both assume conditional independence between variables and allow for the shape of each cluster to vary. The EVI model assumes that the overall volume of each cluster is the same, while this property can also vary for the VVI model.[12] The optimal combination of number of G clusters and model was determined as the one with the highest average approximate Bayesian information criterion over the imputed datasets. Patients were assigned to the cluster with the highest average probability over the imputed datasets fitted with the optimal model and number of G clusters, following relabelling to consistent cluster labels using Stephens' method.[14] Clinical validation was done using summary statistics of the clusters' demography, phenotypic and serological characteristics, clinical diagnosis, and outcome. This was done by restricting each patient to be assigned to only one cluster. Summary statistics were presented as valid percentage, excluding missing data. To account for heterogenous recruitment settings among the source registries we tested differences in cluster assignment between registries using χ^2 tests and visualised these differences using the standardised residuals. Internal validity, and the effect of the multiple imputation, was examined using cluster-wise assessment of cluster stability over the ten imputed datasets using the Jaccard coefficient.[15] We assessed the robustness of the model using leave-one-registry-out analyses, and visualised the results as the empirical cumulative distribution of the log-likelihoods.[16] Due to the lack of an objective reference, no external validation was performed.

We compared the outcome of the clusters with respect to survival and time-to-end-stage kidney disease using a Kaplan–Meier estimator and cumulative incidence function, respectively. The survival analysis evaluated the time from the date of diagnosis (origin and start time) to the date of death or loss to follow-up. The time-to-event analysis for end-stage kidney disease evaluated time from the date of diagnosis (origin and start time) to the date of end-stage kidney disease, date of death, or loss to follow-up, whichever happened first. Patients lost to follow-up were censored at their last known contact date with their respective registry. We built predictive models for survival and kidney survival using Cox proportional hazards models and Fine–Gray models for competing risk, respectively. These models were built to compare the potential predictive value of our subclassification with classification based on clinical diagnosis, ANCA specificity, and a previous cluster-based stratification by Mahr and colleagues.[8] These models were built with gender, age, and source registry as covariates, pooled to account for the uncertainty in cluster

assignment as each patient could potentially be assigned to different clusters in the ten imputed datasets, and model fit was assessed using the Akaike information criterion (AIC; appendix p 21).

To allow for cluster assignment of unseen data, we built a web application using the R package *shiny*.[17] The algorithm gives probabilities for an individual to belong to the clusters, following a simulated expectation-step of the averaged parameters from the multiple imputation runs. To avoid the misclassification of poorly fitting data, the web application flags input with log-odds less than 0 when compared with a uniform noise component.[18] We used R version 4.3.0 for all analyses.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

A total of 3868 patients diagnosed with ANCA-associated vasculitis between Nov 1, 1966, and March 1, 2023, were included in the study across the six registries (Czech Registry n=371, FVSG n=1780, GeVas n=135, POLVAS n=792, RKD n=439, and Skåne Vasculitis Cohort n=351; figure 1). There were 2434 (62·9%) patients with GPA and 1434 (37·1%) with MPA. Mean age at diagnosis was 57·2 years (SD 16·4); 2006 (51·9%) of 3867 patients were men and 1861 (48·1%) were women (table).

Constitutional or musculoskeletal symptoms were present in 2824 (75·5%) of 3738 patients. Kidney involvement was the most common specific organ pattern presentation, seen in 2591 (67·0%) of 3868 patients, followed by lung involvement in 1977 (51·2%%) of 3865 and ear–nose–throat involvement in 1869 (48·3%) of 3867 (table). 1949 (54·0%) of 3607 patients were PR3-ANCA positive, 1344 (37·3%) were MPO-ANCA positive, and 314 (8·7%) were ANCA negative (table). At the time of diagnosis, median serum creatinine was 130·0 µmol/L (IQR 80·0–309·0) and median CRP was 57·0 mg/L (15·0–133·0). In total, 3·2% of data were missing, mostly in the CRP (1059 [27·4%]), serum creatinine (347 [9·0%]), and ANCA status (261 [6·7%]) variables (appendix p 6). Median follow-up was 4·2 years (IQR 1·5–8·7); 642 (16·6%) of 3868 patients reached end-stage kidney disease and 702 (18·1%) patients died. The censoring proportion was 3166 (81·9%) of 3868 for the survival analysis (all lost to follow-up). For the competing risks time-to-event analysis for end-stage kidney disease (n=3076), there were 502 (16·3%) events of end-stage kidney disease and 463 (15·1%) competing deaths, with a censoring proportion of 2111 (68·6%) of 3076 (all lost to follow-up).

Multiple imputation was used to generate ten complete datasets. A total of ten models were fitted (one to five clusters, with two mixture models) to each of the ten complete datasets. The highest average Bayesian information criterion in the ten complete datasets was seen in cluster model VVI with five clusters

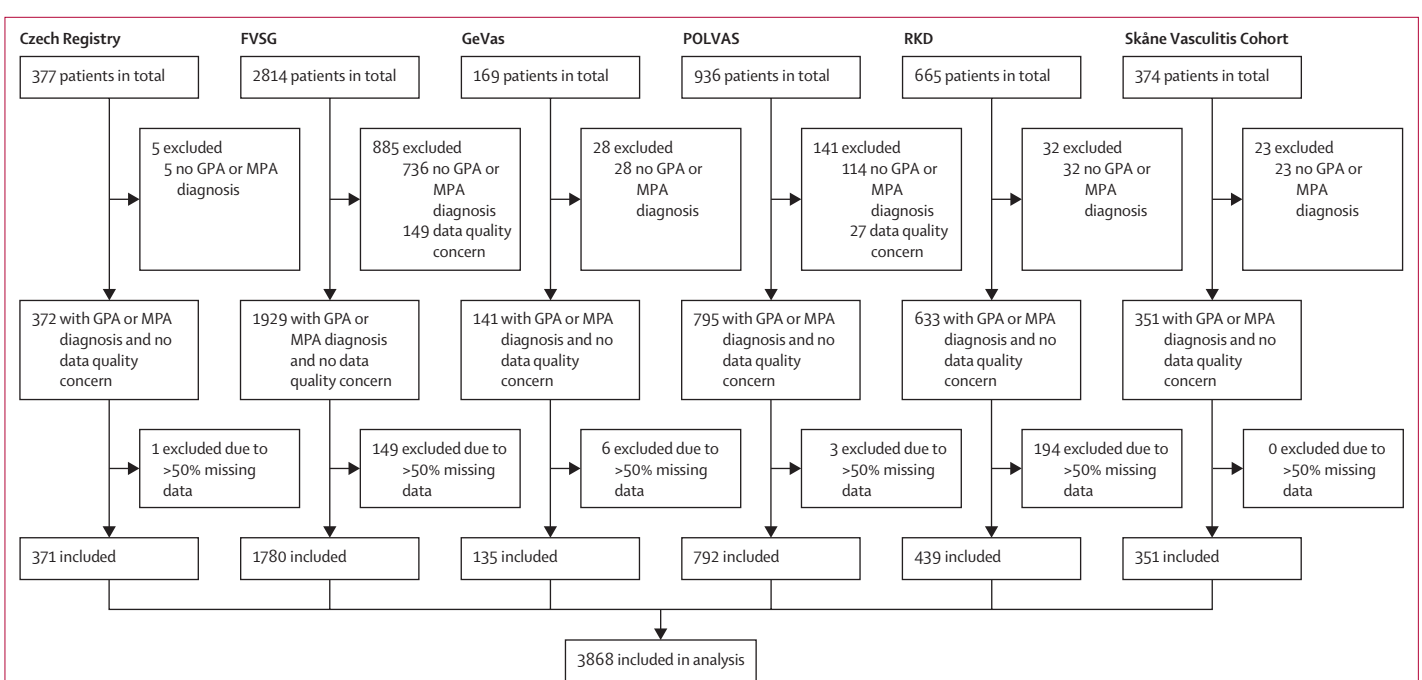


Figure 1: **Study profile**
FVSG=French Vasculitis Study Group Registry. GeVas=Joint Vasculitis Registry in German-speaking Countries. POLVAS=Polish Vasculitis Registry. RKD=Irish Rare Kidney Disease Registry. GPA=granulomatosis with polyangiitis. MPA=microscopic polyangiitis.

	Diagnosis		Cluster				
	GPA (n=2434)	MPA (n=1434)	SK (n=555)	MPO-K (n=782)	PR3-K (n=683)	YR (n=646)	IMS (n=1202)
Demography							
Age, years	53·4 (16·3; n=2413)	63·6 (14·6; n=1422)	66·7 (12·7; n=549)	62·7 (13·8; n=776)	55·9 (15·5; n=679)	47·7 (17·0; n=640)	55·0 (16·2; n=1191)
Gender							
Men	1301/2433 (53·5%)	705 (49·2%)	309 (55·7%)	392 (50·1%)	439 (64·3%)	268 (41·5%)	598/1201 (49·8%)
Women	1132/2433 (46·5%)	729 (50·8%)	246 (44·3%)	390 (49·9%)	244 (35·7%)	378 (58·5%)	603/1201 (50·2%)
Diagnosis							
GPA	2434 (100%)	0	211 (38·0%)	185 (23·7%)	570 (83·5%)	530 (82·0%)	938 (78·0%)
MPA	0	1434 (100%)	344 (62·0%)	597 (76·3%)	113 (16·5%)	116 (18·0%)	264 (22·0%)
Organ pattern							
Constitutional	1534/2384 (64·3%)	740/1356 (54·6%)	323/526 (61·4%)	248/723 (34·3%)	600/676 (88·8)	249/632 (39·4%)	854/1183 (72·2%)
Musculoskeletal	1375/2384 (57·7%)	660/1356 (48·7%)	232/526 (44·1%)	223/723 (30·8%)	541/676 (80·0%)	270/632 (42·7%)	769/1183 (65·0%)
Skin	619/2429 (25·5%)	303 (21·1%)	40 (7·2%)	97 (12·4%)	288 (42·2%)	135/642 (21·0%)	362/1201 (30·1%)
Mucosa	95/2411 (3·9%)	12/1420 (0·8%)	4/551 (0·7%)	5/774 (0·6%)	53/680 (7·8%)	11/634 (1·7%)	34/1192 (2·9%)
Eyes	533/2412 (22·1%)	50/1422 (3·5%)	4/551 (0·7%)	20/775 (2·6%)	173/680 (25·4%)	154/635 (24·3%)	232/1193 (19·4%)
Ear–nose–throat	1697/2433 (69·7%)	172 (12·0%)	111 (20·0%)	110 (14·1%)	477/682 (69·9%)	450 (69·7%)	721 (60·0%)
Lung	1470/2432 (60·4%)	507/1433 (35·4%)	287 (51·7%)	275/781 (35·2%)	456 (66·8%)	271 (42·0%)	688/1200 (57·3%)
Cardiovascular	185/2430 (7·6%)	103/1433 (7·2%)	43/554 (7·8%)	34 (4·3%)	90/682 (13·2%)	37 (5·7%)	84/1199 (7·0%)
Gastrointestinal	170/2429 (7·0%)	103/1431 (7·2%)	33/553 (6·0%)	25/780 (3·2%)	118/682 (17·3%)	17/644 (2·6%)	80/1201 (6·7%)
Kidney	1403 (57·6%)	1188 (82·8%)	554 (99·8%)	770 (98·5%)	676 (99·0%)	147 (22·8%)	444 (36·9%)
Central nervous system	228/2414 (9·4%)	63/1420 (4·4%)	13/552 (2·4%)	15/770 (1·9%)	66/682 (9·7%)	56/636 (8·8%)	141/1194 (11·8%)
Peripheral nervous system	395/2410 (16·4%)	289/1420 (20·4%)	52/552 (9·4%)	32/770 (4·2%)	170/682 (24·9%)	95/634 (15·0%)	335/1192 (28·1%)
Organs involved	4·0 (1·7)	2·9 (1·5)	3·1 (1·2)	2·4 (1·2)	5·4 (1·4)	2·9 (1·5)	4·0 (1·5)
ANCA							
PR3-ANCA positive	1768/2278 (77·6%)	181/1329 (13·6%)	198/532 (37·2%)	131/759 (17·3%)	547/665 (82·3%)	340/567 (60·0%)	733/1084 (67·6%)
MPO-ANCA positive	289/2278 (12·7%)	1055/1329 (79·4%)	307/532 (57·7%)	568/759 (74·8%)	85/665 (12·8%)	126/567 (22·2%)	258/1084 (23·8%)
ANCA negative	221/2278 (9·7%)	93/1329 (7·0%)	27/532 (5·1%)	60/759 (7·9%)	33/665 (5·0%)	101/567 (17·8%)	93/1084 (8·6%)
Laboratory							
Creatinine, µmol/L	96·0 (75·0–216·0; n=2153)	218·0 (104·8–404·8; n=1368)	579·5 (311·3–817·0; n=546)	239·5 (162·0–356·0; n=738)	221·0 (154·0–364·0; n=657)	80·0 (66·0–89·0; n=542)	79·0 (65·0–90·0; n=1038)
CRP, mg/L	64·0 (18·0–142·0; n=1742)	48·0 (13·0–113·0; n=1067)	123·0 (77·0–178·0; n=425)	16·0 (6·0–32·0; n=554)	109·5 (50·0–190·5; n=544)	6·0 (3·0–15·0; n=464)	99·0 (60·0–161·0; n=822)
Cluster							
SK	211 (8·7%)	344 (24·0%)	555 (100%)	0	0	0	0
MPO-K	185 (7·6%)	597 (41·6%)	0	782 (100%)	0	0	0
PR3-K	570 (23·4%)	113 (7·9%)	0	0	683 (100%)	0	0
YR	530 (21·8%)	116 (8·1%)	0	0	0	646 (100%)	0
IMS	938 (38·5%)	264 (18·4%)	0	0	0	0	1202 (100%)
Outcome							
Death	340 (14·0%)	362 (25·2%)	169 (30·5%)	161 (20·6%)	158 (23·1%)	36 (5·6%)	178 (14·8%)
ESKD	299 (12·3%)	343 (23·9%)	231 (41·6%)	221 (28·3%)	136 (19·9%)	11 (1·7%)	43 (3·6%)
Follow-up, years	4·9 (1·8–9·9)	3·2 (1·0–6·8)	2·9 (0·6–5·8)	3·3 (1·4–6·6)	4·3 (1·5–9·0)	6·1 (2·0–11·1)	4·9 (1·8–9·6)

Data are n (%), n/N (%), mean (SD) or mean (SD; n), or median (IQR) or median (IQR; n). Data on ethnicity were not available. GPA=granulomatosis with polyangiitis. MPA=microscopic polyangiitis. SK=severe kidney cluster. MPO-K=anti-MPO kidney involvement cluster. PR3-K=anti-PR3 kidney involvement cluster. YR=young respiratory cluster. IMS=inflammatory multisystem cluster. ANCA=antineutrophil cytoplasmic antibody. CRP=C-reactive protein. ESKD=end-stage kidney disease.

Table: **Baseline phenotypic characteristics and follow-up data within subgroups of GPA and MPA and within each cluster**

(appendix p 14). The patient cluster assignments found in the VVI model with five clusters were retained for further analysis. There was good agreement of patient allocation to clusters in the ten complete datasets (appendix p 9). For descriptive purposes, patients were assigned to the cluster with the highest average probability in the ten complete datasets following a maximum *a posteriori* estimation. The mean cluster assignment

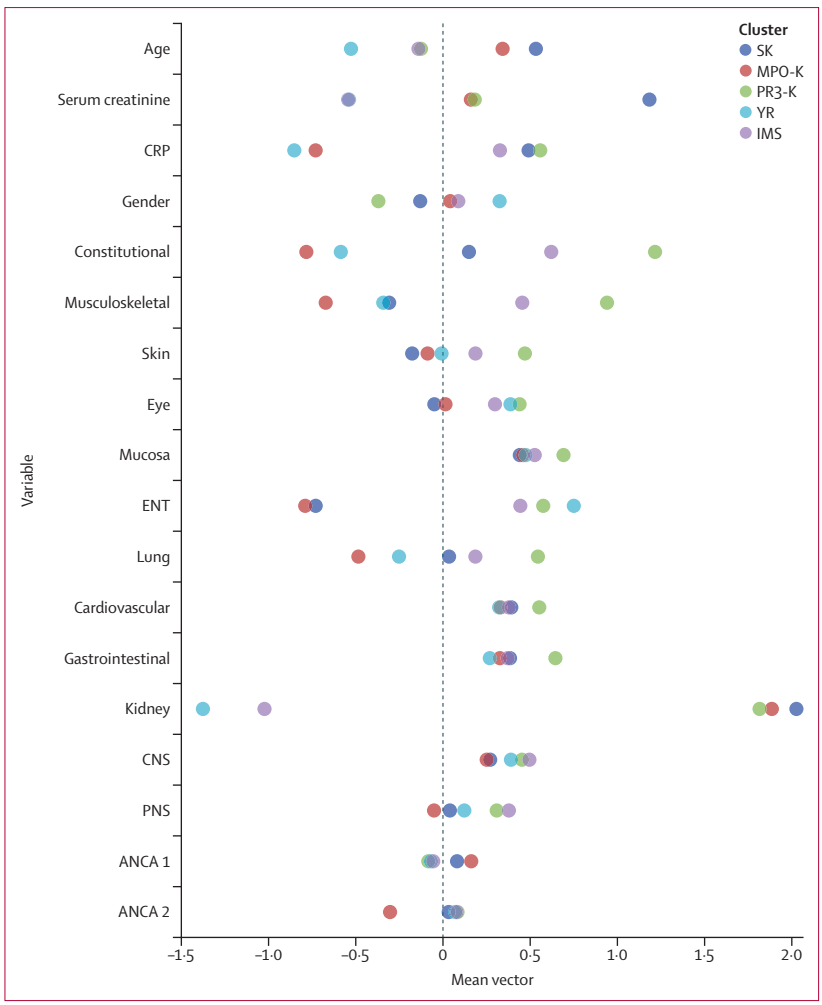


Figure 2: **Mean vectors of each cluster across the dimensions of the combined and observed latent continuous space**

ANCA 1 and ANCA 2 are the one-hot-encoded indicator variables for the three-level nominal variable (ANCA negative, MPO-ANCA positive, and PR3-ANCA positive), and interpreted together they represent the ANCA variable. CRP=C-reactive protein. ENT=ear–nose–throat. CNS=central nervous system. PNS=peripheral nervous system. ANCA=antineutrophil cytoplasmic antibody. SK=severe kidney cluster. MPO-K=anti-MPO kidney involvement cluster. PR3-K=anti-PR3 kidney involvement cluster. YR=young respiratory cluster. IMS=inflammatory multisystem cluster.

probability was 79·4% (appendix p 15). The mean vectors of variables influencing cluster assignment show large separation of clusters based on the presence or absence of kidney involvement, while the separation based on ANCA type is small (figure 2, with a cluster-by-cluster visualisation seen in appendix p 16).

The first cluster, contained 555 (14·3%) of 3868 patients and was characterised by kidney (554 [99·8%] of 555) and lung (287 [51·7%] of 555) involvement, with a very high median serum creatinine (579·5 µmol/L [IQR 311·3–817·0; n=546), and high median CRP (123·0 mg/L [77·0–178·0; n=425]; table). The frequency of ANCA type was relatively evenly distributed between PR3-ANCA (198 [37·2%] of 532) and MPO-ANCA (307 [57·7%] of 532). We named this group the severe kidney (SK) cluster.

The second cluster contained 782 (20·2%) of 3868 patients and was characterised by kidney involvement (770 [98·5%] of 782; median serum creatinine 239·5 µmol/L [IQR 162·0–356·0; n=738]), low-extent organ involvement (mean number of organs involved 2·4 [SD 1·2]), and a low median CRP (16·0 mg/L [IQR 6·0–32·0; n=554]). The cluster was predominantly MPO-ANCA positive (568 [74·8%] of 759). Subsequently, we named this group the anti-MPO with kidney involvement (MPO-K) cluster (table).

The third cluster contained 683 (17·7%) of 3868 patients and was characterised by almost universal kidney involvement (676 [99·0%] of 683) and a median serum creatinine of 221·0 µmol/L [IQR 154·0–364·0; n=657]). This cluster also had wide-extent disease manifestations (mean number of organs involved 5·4 [SD 1·4]) and a high median CRP (109·5 mg/L [IQR 50·0–190·5; n=544]). The cluster was predominantly PR3-ANCA positive (547 [82·3%] of 665) and had a predilection to occur in men (439 [64·3%] of 683). This group was named the anti-PR3 with kidney involvement (PR3-K) cluster.

The fourth cluster contained 646 (16·7%) of 3868 patients who tended to be younger (mean age 47·7 years [SD 17·0; n=640]). It was characterised by a low frequency of kidney involvement (147 [22·8%] of 646), which tended to be mild (median serum creatinine 80·0 µmol/L [IQR 66·0–89·0; n=542]), and the presence of upper respiratory tract (ear–nose–throat; 450 [69·7%] of 646) and lower respiratory tract (lung; 271 [42·0%] of 646) involvement. This cluster was more frequently ANCA negative (101 [17·8%] of 567) than other clusters. We named this group the young respiratory (YR) cluster.

The last cluster was the largest and contained 1202 (31·1%) of 3868 patients. This cluster presented similarly to the PR3-K cluster but with a lower frequency of kidney involvement (444 [36·9%] of 1202) and minimal kidney impairment (median serum creatinine 79·0 µmol/L [IQR 65·0–90·0; n=1038]). Patients in this cluster had a high CRP (median 99·0 mg/L [IQR 60·0–161·0; n=822]) and wider extent organ involvement (mean number of organs involved 4·0 [SD 1·5]). Patients in this cluster were predominantly PR3-ANCA positive (733 [67·6%] of 1084; table). This cluster was named the inflammatory multisystem (IMS) cluster.

The five clusters had distinct outcomes for death and end-stage kidney disease, with the best prognosis for both seen in the YR cluster and the worst prognosis for both seen in the SK cluster (figures 3 and 4). When comparing model fit for prediction of survival and time-to-end-stage kidney disease, our cluster-membership model outperformed models fitted with diagnosis, ANCA status, and an existing cluster stratification regarding both patient survival (ΔAIC 34, 60, and 14) and kidney survival (ΔAIC 349, 366, and 179), minimising the information loss (appendix pp 7–8).

We evaluated cluster stability cluster-wise, with a mean Jaccard index ranging from 0·63 to 0·75 per cluster, indicating stable clusters (appendix p 10). We assessed the robustness of the clusters through six leave-one-registry-out analyses, with stable results (appendix p 17). The cluster assignments were unevenly distributed over the registries, with the most consistent distribution seen for the Swedish Skåne registry (appendix p 18). The FAIRVASC clustering web application for the assignment of previously unseen data to one of the five identified clusters is available online.

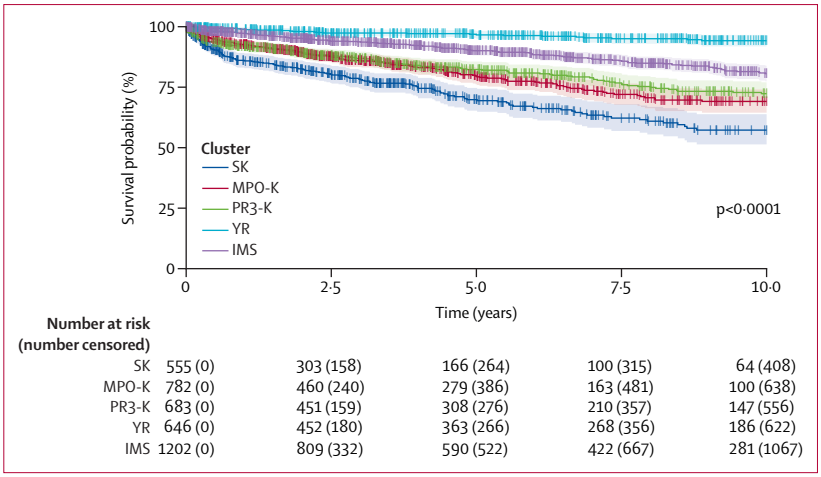
Discussion

In this large study of 3868 patients with ANCA-associated vasculitis, a data-driven subclassification of five distinct clusters had a higher prognostic value than previously established subcategorisations of the disease. We suggest that a subclassification of ANCA-associated vasculitis beyond the binary concepts of PR3-ANCA-positive and MPO-ANCA-positive disease or GPA and MPA should be considered.

Our study supports the results of previous studies using unsupervised learning for the subcategorisation of ANCA-associated vasculitis. Namely, that ANCA-associated vasculitis goes beyond a dichotomous construct. Importantly, Mahr and colleagues[8] subcategorised 673 patients with ANCA-associated vasculitis using hierarchical clustering on principal components to identify five disease subgroups. The three main clusters seen by Mahr and colleagues have since been largely reidentified through latent class analysis,[9] and a unifying subclassification of non-severe, severe PR3-ANCA-associated vasculitis, and severe MPO-ANCA-associated vasculitis has been suggested.[19] Our study largely follows this pattern, with clusters IMS and YR representing a "non-severe ANCA-associated vasculitis" with relatively innocuous disease progression, while clusters PR3-K and MPO-K represent "severe PR3-ANCA-associated vasculitis" and "severe MPO-ANCA-associated vasculitis", respectively. However, our SK cluster assignment seems largely driven by kidney function and high inflammatory response, regardless of ANCA type. Interestingly, ANCA status has limited influence on cluster separation overall. Instead, the main driver in our model is kidney function, an observation already made by Mahr and colleagues.[8] This division of kidney and non-kidney disease subsets is logical, as impaired kidney function has been firmly established as a key determinant of unfavourable outcomes in patients with ANCA-associated vasculitis.[20]

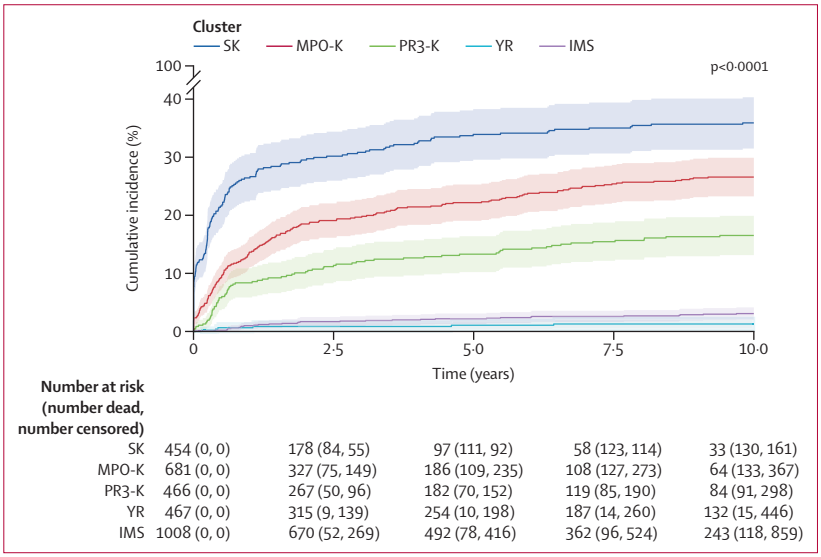
The five identified clusters have distinct phenotypic and biochemical presentation, as well as disease outcome. However, linking the five-cluster stratification to defining genetic, epigenetic, or other pathogenic precursors is hampered by the paucity of studies stratified by phenotypic expression. Naturally, investigations into the pathogenesis of ANCA-associated vasculitis have been largely focused on the disease groups or ANCA pattern.[21] The subclassification we present highlights the need for studies on the role of inter-organ vascular diversity, systemic inflammatory response, and ageing in phenotypic expression.

The use of data-driven phenotypic subclassifications of disease is gaining traction in the heterogenous disease setting of systemic rheumatic disease.[22,23] Several studies have been performed in ANCA-associated vasculitis, but clustering attempts have also been made within the ANCA-associated vasculitis disease subtypes.[24–26] However, studies are lacking in both sample size and information granularity.



Number at risk (number censored)					
SK	555 (0)	303 (158)	166 (264)	100 (315)	64 (408)
MPO-K	782 (0)	460 (240)	279 (386)	163 (481)	100 (638)
PR3-K	683 (0)	451 (159)	308 (276)	210 (357)	147 (556)
YR	646 (0)	452 (180)	363 (266)	268 (356)	186 (622)
IMS	1202 (0)	809 (332)	590 (522)	422 (667)	281 (1067)

Figure 3: **Kaplan–Meier plot for 10-year survival**
P value obtained by the log-rank test. Shaded areas show 95% CIs. SK=severe kidney cluster. MPO-K=anti-MPO kidney involvement cluster. PR3-K=anti-PR3 kidney involvement cluster. YR=young respiratory cluster. IMS=inflammatory multisystem cluster.



Number at risk (number dead, number censored)					
SK	454 (0, 0)	178 (84, 55)	97 (111, 92)	58 (123, 114)	33 (130, 161)
MPO-K	681 (0, 0)	327 (75, 149)	186 (109, 235)	108 (127, 273)	64 (133, 367)
PR3-K	466 (0, 0)	267 (50, 96)	182 (70, 152)	119 (85, 190)	84 (91, 298)
YR	467 (0, 0)	315 (9, 139)	254 (10, 198)	187 (14, 260)	132 (15, 446)
IMS	1008 (0, 0)	670 (52, 269)	492 (78, 416)	362 (96, 524)	243 (118, 859)

Figure 4: **Cumulative incidence of end-stage kidney disease over 10 years with death as a competing risk**
P value obtained by Gray's test. Shaded areas show 95% CIs. SK=severe kidney cluster. MPO-K=anti-MPO kidney involvement cluster. PR3-K=anti-PR3 kidney involvement cluster. YR=young respiratory cluster. IMS=inflammatory multisystem cluster.

For the **FAIRVASC clustering web application** see https://fairvasc.shinyapps.io/clusteranalysis/

Our study is based on the largest cohort yet to be described in ANCA-associated vasculitis.[10] Considering the data-hungriness of machine learning models, the sample size is a major strength of this study. This being an exploratory analysis using unsupervised learning, with no objective reference, we performed no external validation of the cluster model. Instead, we have provided detailed checks using repeatedly held-out data, which show support for the model. However, the large study cohort is based on real-world patient data from six well established European ANCA-associated vasculitis registries and also captures patients with ANCA-associated vasculitis not meeting the inclusion criteria of clinical trials. This is a considerable strength and speaks to the generalisability of our results, and further highlights the need for inclusion of this patient group in clinical trials. Although the registries are known to have heterogenous recruitment settings, this has been accounted for in the imputation model. The even distribution of individuals from the population-based Swedish Skåne registry compared with the other five FAIRVASC registries and the results of the six leave-one-registry-out validations are further signs of the generalisability of the clusters. As such, this stratification model could provide enhanced sample homogeneity in mechanistic studies, novel compound development, and clinical trials.

However, there are limitations to this study. The assembly of this large cohort has meant trade-offs between sample size and granularity. Our model lacks data of evident interest in the subclassification of disease, such as biomarkers of disease activity, radiography, and histology. The lack of radiographic data is especially troublesome considering the increased understanding of pulmonary fibrosis as an important clinical feature in MPA.[27] Furthermore, the combination of immunoassay and immunofluorescent ANCA in this study is a limitation. However, this was done by necessity (part of the cohort was established before the widespread use of immunoassay), is in line with current classification methods,[28,29] and extends the generalisability to communities where immunoassay is not widely available. We do, however, lack sufficient data to explore additional interesting aspects such as the differences in treatment effect per cluster, outcome in terms of relapse rate, or patient-reported outcomes. Another limitation is the absence of data regarding ethnicity. Such data were not recorded by four of the source registries and difficult to appropriately harmonise between the remaining two. Furthermore, to ensure the quality of the imputation of missing values and the following analysis, we decided to exclude patients where the completeness of data was deemed too low. It should, however, be noted that such a decision is subject to selection bias, and as such, is an important limitation to this study. It should also be noted that the prediction models were not built for the purpose of causal inference but to allow for comparison of the potential utility of different methods of subclassification for predicting key outcomes in ANCA-associated vasculitis. For causal purposes the models are flawed due to the inherent selection bias of hazard ratios, the potential for unmeasured confounding, and for the competing risks analysis the choice of using Fine–Gray models as these estimate the total and not only direct effect of the exposure.[30]

Lastly, our model is sufficiently complex to not be easily interpretable. This prohibits the use of a human-evaluable decision tree or risk score for the classification of disease. Instead, we opted for a publicly available web application for the assignment of new patients. In addition, to give probabilities for assignment to the different clusters, the web application checks the validity of the input data using a robustness measure devised by Coretto and Hennig.[18]

These limitations raise interesting questions regarding how this subclassification could be used in clinical practice. Further studies are needed to assess the generalisability of the model, especially to other geographical regions, but also how the phenotypic expression and cluster affiliation evolve before and after the date of diagnosis. Studies providing data regarding management and treatment are needed to investigate how cluster assignment should affect management decisions. This study could provide a new paradigm for stratification of ANCA-associated vasculitis, but the data-driven methodology is also expandable for the subclassification of other systemic inflammatory and autoimmune disorders with complex phenotypic expressions.

To conclude, this study further reinforces the hypothesis that ANCA-associated vasculitis is not merely a binary construct. We suggest a data-driven subclassification of ANCA-associated vasculitis into five distinct groups identified through model-based clustering of 3868 patients. This subclassification, based on real-world patient data, could provide more accurate patient stratification for therapeutic, epidemiological, and basic research in ANCA-associated vasculitis.

Contributors

KG, AW, MS, MAL, and AJM conceived and designed the study. The FAIRVASC consortium collected the data. KG and AJM directly accessed and verified the underlying data. KG, AW, LA, JNa, and JNg performed the data analysis and statistical calculation. KG, ZH, PL, JM, DOS, XP, MR, MS, BT, VT, MT, AV, KW, MAL, and AJM interpreted the scientific data. KG drafted the manuscript, which was revised by all authors. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

PL has received grants or contracts from the German Research Society, Federal Ministry of Education and Research, German Society for Rheumatology, John Grube Foundation, and Vifor Pharma; consulting fees from AbbVie, AstraZeneca, GSK, Novartis, UCB, and Vifor Pharma; payment or honoraria for lectures, presentations, speakers, bureaus, manuscript writing, or educational events from AstraZeneca, Boehringer Ingelheim, GSK, Janssen, Novartis, Rheumaakademie, UCB, and Vifor Pharma; support for attending meetings or travel from Vifor Pharma; and has participated on Data Safety Monitoring Boards or Advisory Boards of AbbVie, AstraZeneca, GSK, and Vifor Pharma. XP has

received grants as an investigator of academic studies of ANCA-associated vasculitis for which rituximab was provided by Roche Pharma. MS has received consulting fees from Vifor Pharma, Otsuka, and Hansa Biopharma. BT has received consulting fees from Novartis, AstraZeneca, CSL Vifor, GSK, and Boehringer Ingelheim; payment or honoraria for lectures, presentations, speakers, bureaus, manuscript writing, or educational events from Novartis, AstraZeneca, CSL Vifor, GSK, and Boehringer Ingelheim; support for attending meetings or travel from GSK, AstraZeneca, and Novartis; and participates on the Advisory Board for Amgen. KW is a member of the Advisory Board at CSL Vifor. All other authors declare no competing interests.

Data sharing

Supporting data are not publicly available due to ethical, legal, and commercial reasons. Statistical analysis and code are available at https://github.com/karlgi/clusteranalysis.

Acknowledgments

We would like to thank all the patient contributors with lived experience of ANCA-associated vasculitis who have been involved in this study, and especially Julie Power for drafting the statement in the manuscript regarding this involvement. This work is generated within the European Reference Network for Rare Immunodeficiency, Autoinflammatory and Autoimmune Diseases (ERN-RITA) and funded by the European Union's Horizon 2020 research and innovation programme under the European Joint Programme on Rare Diseases (EJP RD; COFUND-EJP 825575). Additional funders are: Crafoordska stiftelsen (20220623; to AJM), Vetenskapsrådet (2019-00263; to AJM), the Science Foundation Ireland (13/RC/2106_P2 and 11/Y/B2093; to ML), and the European Union's Horizon 2020 research and innovation programme under the EJP RD (EJP-RD/I/FAIRVASC/03/2020 [K/NCB/000058]; to JM and KW).

References

1 Domínguez-Quintana M, Alba MA, Hinojosa-Azaola A. Classification of ANCA-associated vasculitis: differences based on ANCA specificity and clinicopathologic phenotype. *Rheumatol Int* 2021; **41:** 1717–28.

2 Jennette JC, Falk RJ, Bacon PA, et al. 2012 revised International Chapel Hill Consensus Conference nomenclature of vasculitides. *Arthritis Rheum* 2013; **65:** 1–11.

3 Kitching AR, Anders HJ, Basu N, et al. ANCA-associated vasculitis. *Nat Rev Dis Primers* 2020; **6:** 71.

4 Lyons PA, Rayner TF, Trivedi S, et al. Genetically distinct subsets within ANCA-associated vasculitis. *N Engl J Med* 2012; **367:** 214–23.

5 Lionaki S, Blyth ER, Hogan SL, et al. Classification of antineutrophil cytoplasmic autoantibody vasculitides: the role of antineutrophil cytoplasmic autoantibody specificity for myeloperoxidase or proteinase 3 in disease recognition and prognosis. *Arthritis Rheum* 2012; **64:** 3452–62.

6 Unizony S, Villarreal M, Miloslavsky EM, et al. Clinical outcomes of treatment of anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis based on ANCA type. *Ann Rheum Dis* 2016; **75:** 1166–69.

7 Watanabe H, Sada KE, Harigai M, et al. Exploratory classification of clinical phenotypes in Japanese patients with antineutrophil cytoplasmic antibody-associated vasculitis using cluster analysis. *Sci Rep* 2021; **11:** 5223.

8 Mahr A, Katsahian S, Varet H, et al. Revisiting the classification of clinical phenotypes of anti-neutrophil cytoplasmic antibody-associated vasculitis: a cluster analysis. *Ann Rheum Dis* 2013; **72:** 1003–10.

9 Wójcik K, Biedroń G, Wawrzycka-Adamczyk K, et al. Subphenotypes of ANCA-associated vasculitis identified by latent class analysis. *Clin Exp Rheumatol* 2021; **39** (suppl 129)**:** 62–68.

10 Gisslander K, Rutherford M, Aslett L, et al. Data quality and patient characteristics in European ANCA-associated vasculitis registries: data retrieval by federated querying. *Ann Rheum Dis* 2023; **83:** 112–20.

11 McGlinn K, Rutherford MA, Gisslander K, et al. FAIRVASC: a semantic web approach to rare disease registry integration. *Comput Biol Med* 2022; **145:** 105313.

12 McParland D, Gormley I. Model based clustering for mixed data: clustMD. *Adv Data Anal Classif* 2015; **10:** 155–69.

13 Basagaña X, Barrera-Gómez J, Benet M, Antó JM, Garcia-Aymerich J. A framework for multiple imputation in cluster analysis. *Am J Epidemiol* 2013; **177:** 718–25.

14 Stephens M. Dealing with label switching in mixture models. *J R Stat Soc Series B Stat Methodol* 2002; **62:** 795–809.

15 Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal* 2007; **52:** 258–71.

16 O'Hagan A, Murphy TB, Gormley IC, McNicholas PD, Karlis D. Clustering with the multivariate normal inverse Gaussian distribution. *Comput Stat Data Anal* 2016; **93:** 18–30.

17 Chang W CJ, Allaire J, Sievert C, et al. Shiny: web application framework for R. 2023. https://cran.r-project.org/package=shiny (accessed July 22, 2024).

18 Coretto P, Hennig C. A simulation study to compare robust clustering methods based on mixtures. *Adv Data Anal Classif* 2010; **4:** 111–35.

19 Mahr A, Specks U, Jayne D. Subclassifying ANCA-associated vasculitis: a unifying view of disease spectrum. *Rheumatology (Oxford)* 2019; **58:** 1707–09.

20 Flossmann O, Berden A, de Groot K, et al. Long-term patient survival in ANCA-associated vasculitis. *Ann Rheum Dis* 2011; **70:** 488–94.

21 Nakazawa D, Masuda S, Tomaru U, Ishizu A. Pathogenesis and therapeutic interventions for ANCA-associated vasculitis. *Nat Rev Rheumatol* 2019; **15:** 91–101.

22 Sobanski V, Giovannelli J, Allanore Y, et al. Phenotypes determined by cluster analysis and their survival in the prospective European Scleroderma Trials and Research Cohort of patients with systemic sclerosis. *Arthritis Rheumatol* 2019; **71:** 1553–70.

23 McCoy SS, Woodham M, Bartels CM, et al. Symptom-based cluster analysis categorizes Sjögren's disease subtypes: an international cohort study highlighting disease severity and treatment discordance. *Arthritis Rheumatol* 2022; **74:** 1569–79.

24 Wang J, Li R, Zhou W, et al. Clinical phenotypes and prognoses of microscopic polyangiitis based on kidney biopsies. *Arthritis Res Ther* 2023; **25:** 239.

25 Thayakaran R, Goel R, Adderley NJ, et al. Cluster analysis of patients with granulomatosis with polyangiitis (GPA) based on clinical presentation symptoms: a UK population-based cohort study. *Arthritis Res Ther* 2022; **24:** 201.

26 Rubenstein E, Maldini C, Vaglio A, et al. Cluster analysis to explore clinical subphenotypes of eosinophilic granulomatosis with polyangiitis (Churg-Strauss). *J Rheumatol* 2023; **50:** 1446–53.

27 Maillet T, Goletto T, Beltramo G, et al. Usual interstitial pneumonia in ANCA-associated vasculitis: a poor prognostic factor. *J Autoimmun* 2020; **106:** 102338.

28 Robson JC, Grayson PC, Ponte C, et al. 2022 American College of Rheumatology/European Alliance of Associations for Rheumatology classification criteria for granulomatosis with polyangiitis. *Ann Rheum Dis* 2022; **81:** 315–20.

29 Suppiah R, Robson JC, Grayson PC, et al. 2022 American College of Rheumatology/European Alliance of Associations for Rheumatology classification criteria for microscopic polyangiitis. *Ann Rheum Dis* 2022; **81:** 321–26.

30 Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med* 2020; **39:** 1199–236.