Clinical phenotypes in α_1-antitrypsin deficiency: a cluster analysis on EARCO data

Anse Somers [1,24], Kenneth Verstraete [1,2,24], Ella Demesmaeker[3], María Torres-Duran[4,5], Alice M. Turner [6,7], Hanan Tanash[8], Carlota Rodríguez-García[9], Jens-Ulrik Stæhr Jensen [10,11], Angelo Guido Corsico[12,13], José Luis López-Campos [14,15], Kenneth R. Chapman [16], Christian F. Clarenbach [17], Catarina Guimarães[18], Eva Bartošovská[19], José María Hernández-Pérez [20], Juan Luis Rodriguez-Hermosa [21], Beatriz Lara [22], Marc Miravitlles [23], Wim Janssens [1,3] and Stephanie Everaerts [1,3]

[1]Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), Department of Chronic Diseases and Metabolism (CHROMETA), KU Leuven, Leuven, Belgium. [2]STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium. [3]Clinical Department of Respiratory Diseases, University Hospitals Leuven, Leuven, Belgium. [4]Servicio de Neumología, Hospital Álvaro Cunqueiro. NeumoVigo I+i Research Group, IIS Galicia Sur, Vigo, Spain. [5]Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Madrid, Spain. [6]Respiratory Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [7]School of Applied Health Sciences, University of Birmingham, Birmingham, UK. [8]Department of Respiratory Medicine and Allergology, Skåne University Hospital, Lund University, Malmö, Sweden. [9]Servicio de Neumología, Complejo Hospitalario Clínico-Universitario de Santiago, Santiago de Compostela, Spain. [10]Section of Respiratory Medicine, Department of Medicine, Herlev and Gentofte Hospital, University of Copenhagen, Hellerup, Denmark. [11]Department of Clinical Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [12]Department of Internal Medicine and Medical Therapeutics, University of Pavia, Pavia, Italy. [13]Unit of Respiratory Diseases, IRCCS Policlinico San Matteo Foundation, Pavia, Italy. [14]Unidad Médico-Quirúrgica de Enfermedades Respiratorias. Instituto de Biomedicina de Sevilla (IBiS), Hospital Universitario Virgen del Rocío/Universidad de Sevilla, Seville, Spain. [15]Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES). Instituto de Salud Carlos III, Madrid, Spain. [16]Asthma and Airway Centre, University Health Network, Toronto Western Hospital, and Division of Respiratory Medicine, Department of Medicine, University of Toronto, Toronto, ON, Canada. [17]Division of Pulmonology, University Hospital Zurich, Zurich, Switzerland. [18]Pulmonology Department, Hospital da Senhora da Oliveira, Unidade Local de Saúde Alto Ave, Guimarães, Portugal. [19]Department of Pneumology, Thomayer Hospital, First Faculty of Medicine, Charles University, Prague, Czech Republic. [20]Pneumology Department, Hospital Universitario Nuestra Señora de La Candelaria, Santa Cruz de Tenerife, Spain. [21]Pulmonology Department. Hospital Clínico San Carlos, Department of Medicine, School of Medicine, Universidad Complutense de Madrid, Instituto de Investigación Sanitaria del Hospital Clínico San Carlos (IdISSC), School of Medicine, Universidad Antonio de Nebrija, Madrid, Spain. [22]Department of Respiratory Medicine, University Hospitals of Coventry and Warwickshire, Coventry, UK. [23]Department of Pneumology, Hospital Universitari Vall d'Hebron/Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Health Care Provider of the European Reference Network on Rare Respiratory Diseases (ERN LUNG), Barcelona, Spain. [24]These authors contributed equally.

Corresponding author: Stephanie Everaerts (stephanie.everaerts@uzleuven.be)

Shareable abstract (@ERSpublications)
Six clusters of AATD clinical phenotypes – mainly differing in age at diagnosis, lung function and smoking, in addition to genotype – were identified using EARCO baseline data, and hold promise for therapeutic and prognostic value https://bit.ly/4swWLkY

Cite this article as: Somers A, Verstraete K, Demesmaeker E, *et al*. Clinical phenotypes in α_1-antitrypsin deficiency: a cluster analysis on EARCO data. *ERJ Open Res* 2026; 12: 01597-2025 [DOI: 10.1183/23120541.01597-2025].

Abstract

Background α-1 antitrypsin deficiency (AATD) is a rare, monogenic disorder predisposing individuals to liver and lung diseases. Yet, even within the same genotype, AATD is very heterogeneous in clinical presentation. This heterogeneity suggests a complex interplay of environmental and (epi)genetic factors, highlighting the need for a detection and study of clinical phenotypes.

Methods Here, we performed a cluster analysis using baseline data from the European Alpha-1 Research Collaboration (EARCO) to identify distinct clinical phenotypes in AATD.

Results We identified six clusters of AATD clinical phenotypes using K-prototypes with a cross-validated weighted F1 score of 0.926. Moreover, using a trained random forest classifier, we identified age at

Received: 14 Nov 2025
Accepted: 23 Dec 2025

diagnosis, lung function and tobacco consumption as the most important features in distinguishing these clusters, in addition to the AATD-associated genotype.

Conclusions Overall, these findings underscore the fact that, beyond genotype, a multitude of other factors significantly influence AATD clinical phenotypes. Accordingly, this emphasises the importance of integrating (epi)genetic, environmental and behavioural data in future fundamental and translational research to unravel the complex mechanisms underlying patient heterogeneity. Moreover, longitudinal follow-up of individuals allocated to these clusters will help to better understand disease progression and to assess the potential benefits of specific therapies in different phenotypes.

Introduction

α-1 antitrypsin (AAT) deficiency (AATD) is a rare, monogenic disorder caused by mutations in *SERPINA1*, the gene encoding AAT, predisposing individuals to the development of liver and lung diseases [1–3]. AAT is a key serum protease inhibitor with immunomodulatory functions and is produced primarily in the liver, though there is also extrahepatic expression, including in the lungs [4–11]. Deficient AAT levels allow lung parenchymal damage due to uncontrolled neutrophil elastase activity, instigating early-onset emphysema [10, 12–15]. Additionally, the most common mutations cause misfolded AAT to accumulate and polymerise, provoking liver damage and further immune activation in the lungs, thereby aggravating lung parenchymal damage [16–18]. Notably, smoking greatly accelerates the onset and severity of lung parenchymal damage [17, 19–21].

Although a monogenic disorder, over 100 AAT variants have been described, with the most prevalent being the normal or wild-type M variant, and the deficiency-associated S and Z variants [22–27]. Yet, even within the same genotype, AATD is a very heterogeneous condition regarding clinical presentation. This heterogeneity suggests a complex interplay of individual-specific environmental and (epi)genetic factors, highlighting the need for a detection and study of clinical phenotypes. Here, we performed a cluster analysis using baseline data from the European Alpha-1 Research Collaboration (EARCO) to identify distinct clinical phenotypes in AATD. Overall, we aimed to organise the heterogeneous group of individuals with AATD into subgroups based on their degree of similarity by employing a multivariate algorithm, considering that a better stratification of patients may hold promise for more personalised treatment and prognosis in the future [28, 29].

Material and methods

Study population

We used data from the EARCO registry, a multinational observational cohort study enrolling individuals diagnosed with AATD, as confirmed by AAT serum level and/or genotyping [30]. We included all registered individuals from the inception of the registry in February 2020 up to 13 February 2024. Only baseline data were used due to the paucity of longitudinal follow-up data at the time of analysis. A total of 2540 individuals with 1110 recorded parameters were eligible for analysis.

Missing data and pre-processing

A preliminary missing data analysis was conducted to assess data completeness and to identify the subset of patients suitable for clustering. The clustering was performed as a complete case analysis, hence only subjects with non-missing values for the selected features were included. Some pre-processing steps are mentioned in supplementary data S1. Parameters with zero variance or with >80% missing data were excluded. Final feature selection was based on expert clinical judgment (WJ and SE).

Selected features

The final analysis included both continuous and categorical variables:

- Continuous variables: AAT level at diagnosis, age at diagnosis, body mass index (BMI), pack-years, forced expiratory volume in 1 s (FEV_1) relative to predicted value, diffusion capacity of the lung for carbon monoxide (D_{LCO}) relative to predicted value, COPD assessment test (CAT) score, neutrophils and eosinophils.
- Categorical variables: sex, smoking history, index case, augmented therapy, COPD, emphysema, chronic bronchitis, bronchiectasis, asthma, lung cancer, other lung diseases, dyspnoea, cough, sputum, asymptomatic status, fatty liver, cirrhosis, liver fibrosis, deranged liver function tests, hepatocellular carcinoma, jaundice, neonatal jaundice, other liver disease and other AATD-related diseases.

Clustering methodology

To accommodate both continuous and categorical data types, we employed the K-prototypes algorithm, an extension of K-means that integrates elements of the K-modes algorithm to handle categorical variables [31].

The optimal number of clusters was determined using a supervised machine-learning model. Due to the tabular nature of the data, random forest was the model of choice [32].

Cluster number selection

We repeated the following procedure for M=2–10 clusters:

1. Apply K-prototypes on the baseline features to partition the dataset into M clusters.
   The continuous and categorical features were used as input for the K-prototypes algorithm together with parameter M to find M clusters based on the provided features.
2. Use cluster assignments as classification labels. Each subject in the dataset was assigned to a cluster by the K-prototypes algorithm through cluster labels. The possible cluster labels ranged from 1 to M.
3. Train a random forest classifier using baseline features to predict these cluster labels.
   The continuous and categorical features were used as input to train a random forest classification model that predicts the cluster labels of subjects based on their features.
4. Perform 10-fold cross-validation, repeated 100 times, stratified by cluster assignment.
   To avoid evaluating the trained model on subjects on which it was trained, we used the cross-validation method. The 10-fold cross-validation method randomly splits the dataset into 10 subsets (folds). The model is then trained on nine of the 10 folds and evaluated on the remaining fold. This process is repeated such that each fold is used as an evaluation fold once. On each evaluation fold, the F1 score is calculated [33]. This score is an evaluation metric measuring a balance (harmonic mean) between positive predictive value and sensitivity and ranges between 0 (worst value) and 1 (best value). The 10-fold cross-validation process was repeated 100 times over different random splits.
5. Calculate the mean F1 score across all repetitions and folds. The mean F1 score is calculated over the 100 different random splits. As such, for each number of clusters, a mean F1 score is calculated [33]. The number of clusters yielding the highest mean F1 score was selected.

Feature importance and reduction

To estimate the contribution of each feature to cluster discrimination, we used the trained random forest models during each iteration at step four to compute feature importance scores. This approach quantified how informative each feature is in distinguishing between the clusters. After ranking the features according to their importance, the number of features was reduced based on the importance and ease of collection.

Statistical analysis

We performed descriptive statistics for each cluster across domains including demographics, lung function, blood biomarkers, medical history and status, and liver function tests. Continuous variables are presented as mean±SD or median with interquartile range (Q1–Q3), as appropriate. Categorical variables are reported as n (%). To assess overall differences between clusters, a one-way ANOVA was used for continuous variables. When significant, pairwise comparisons were conducted using *post hoc* t-tests with Bonferroni–Holm correction to adjust for multiple testing. Python 3.8 (Python Software Foundation) and the SciPy (v.1.10.1), statsmodels (v.0.14.1) and kmodes (v.0.12.2) packages were used. The scikit-learn (v.1.3.2) package was used to perform supervised machine learning.

Results

Initially, 1196 individuals from the 2540 individuals in the EARCO registry (data extracted on 13 February 2024) were used with 34 baseline variables. After performing the supervised machine-learning method, the determined optimal number of clusters was six, and the number of baseline variables was reduced from 34 to 15. The number of subjects for the final analysis after re-adding the 42 individuals who had missing data for the omitted 19 baseline variables was therefore 1238. The full flowchart is visualised in figure 1. Based on the highest cross-validated-weighted F1 score (0.926), the optimal number of clusters was determined to be six (figure 2). Cluster characteristics are described in table 1, sorted by decreasing feature importance as determined by the trained random forest classifier (figure 3). The six most distinguishing continuous variables (age at AATD diagnosis, FEV_1, D_{LCO}, pack-years, CAT score and AAT level) are also shown in the radar plot (figure 4). All statistically significant differences between clusters are illustrated in figure 5. Altogether, we identified six clusters, each representing a distinct AATD clinical phenotype.

Overall, individuals in cluster 2 were significantly younger at their AATD diagnosis, whereas individuals in cluster 4 were significantly older compared with all other clusters. Regarding lung function measurements (FEV_1 and D_{LCO}), clusters 1, 2 and 3 grouped individuals with normal lung function. In contrast, clusters 4, 5 and 6 grouped individuals with COPD diagnosis based on low lung function and high CAT scores, with most individuals suffering from lung emphysema and dyspnoea. The FEV_1 differed significantly between these three clusters, being the lowest in cluster 6. Similarly, the D_{LCO} was

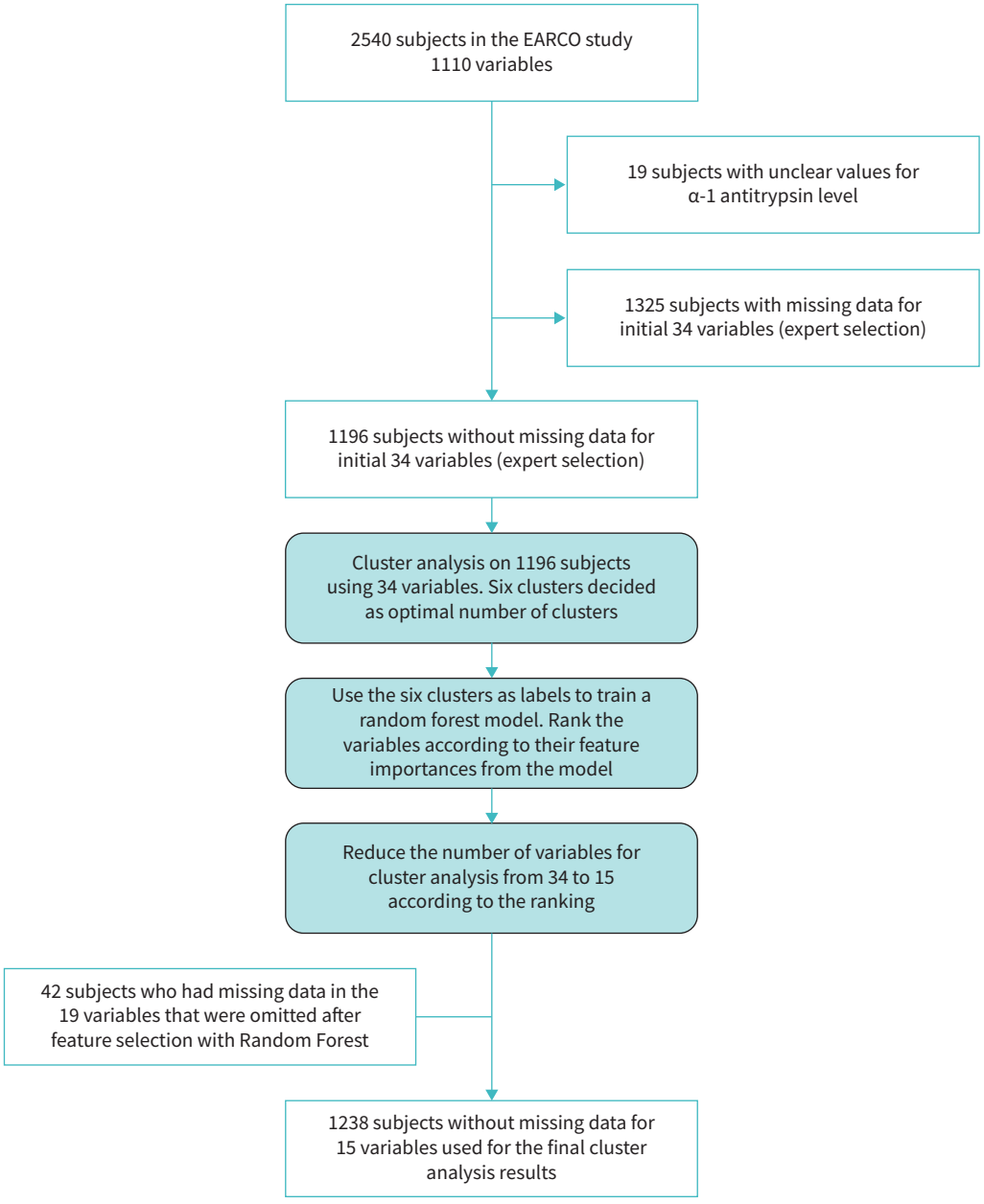

FIGURE 1 Flowchart regarding the selected study population of individuals with α-1 antitrypsin deficiency (AATD) in the European Alpha-1 Research Collaboration (EARCO) registry.

significantly lowest in cluster 6. Notably, the pack-years significantly differed between all clusters, except between clusters 1 and 2, and between clusters 3 and 5. All clusters grouped individuals of various AATD-associated genotypes, but shifted to a dominant ZZ genotype in clusters 5 and 6, coinciding with the lowest serum AAT levels of all clusters. Notably, in all clusters, most individuals were identified as index cases. The proportion of index cases was lowest in cluster 2 and highest in cluster 4, with significant differences in proportion between clusters 1, 2 and 3 *versus* clusters 4, 5 and 6, and between clusters 4 and 5. For the remaining variables with lower differentiating value, only serum neutrophil count did differ significantly between all clusters, except between clusters 1, 2 and 3.

Discussion

In this cluster analysis, we identified six AATD clinical phenotypes mainly differentiated by age at AATD diagnosis, lung function and tobacco consumption, in addition to the AATD-associated genotype.

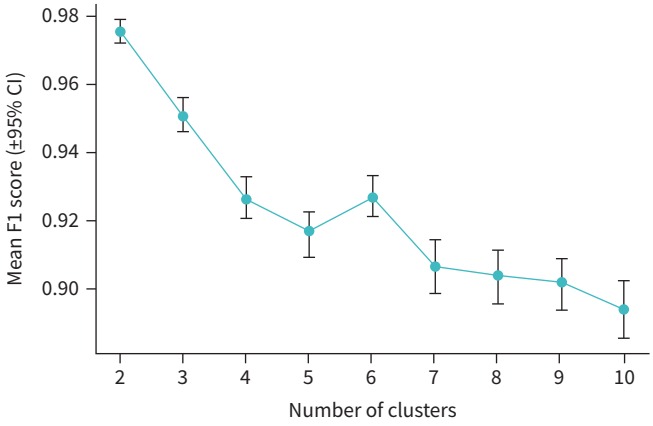

FIGURE 2 Random forest classification of cluster assignments. The weighted F1 score accounts for class imbalance and combines precision and recall into a single performance metric. The number of clusters yielding the highest mean weighted F1 score was selected.

Age at AATD diagnosis has the largest differentiating power and may be reflected by changes in lifestyle, specifically regarding smoking behaviour, after receiving this diagnosis. In particular, individuals from cluster 2 who received their AATD diagnosis at the youngest age, report a low tobacco consumption, have a normal lung function at inclusion in the registry and the majority do not suffer from emphysema nor COPD. Moreover, their clinical presentation resembled individuals in cluster 1, consisting of individuals who have never smoked. In contrast, individuals in cluster 4 who received their AATD diagnosis at the oldest age, report the highest tobacco consumption. Interestingly, clusters 2 and 4 also had the lowest and highest number of index cases respectively. Hence, this highlights the importance of early AATD diagnosis, including *via* genetic screening, and awareness regarding the impact of smoking, corroborating the findings of the longitudinal follow-up from individuals of the Swedish National AATD Registry [34–36].

While individuals in clusters 3 and 6 did not have a significant different age at diagnosis and only a limited difference in pack-years, individuals in cluster 6 presented with a severely reduced lung function whereas individuals in cluster 3 presented with a normal lung function. This discrepancy in lung function could be attributed to the AATD-associated genotype, being predominantly SZ in cluster 3 and ZZ in

TABLE 1 Six clusters of clinical phenotypes in α-1 antitrypsin deficiency

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Count n	263	193	209	141	194	238
Age at diagnosis, years	52.1±11.4	17.5±11.8	48.6±12.5	62.4±11.0	50.0±11.8	47.2±13.4
FEV_1, % pred	104.0±18.1	100.2±14.8	101.5±17.8	58.8±24.3	66.4±20.5	38.4±15.1
Pack-years	0.0±0.0	0.4±1.7	17.2±14.5	46.0±29.3	13.7±13.3	19.1±15.5
Smoking history: yes/no, %	0/263 (0%)	21/172 (11%)	208/1 (100%)	138/3 (98%)	146/48 (75%)	199/39 (84%)
CAT score	5.5±6.2	3.1±4.8	4.4±5.3	11.8±8.2	7.8±5.3	19.6±6.9
AAT level, $mg \cdot dL^{-1}$	47.9±20.8	44.2±21.6	58.8±20.2	71.3±21.4	24.2±10.0	26.2±13.9
D_{LCO}, % pred	90.6±19.2	93.1±17.4	88.8±19.1	57.0±23.1	60.4±22.7	42.6±15.5
SS/SZ/ZZ/other genotype, %	8/52/29/12	6/44/41/9	18/61/9/12	26/62/3/10	0/2/86/12	0/3/86/11
COPD: yes/no, %	7/256 (3%)	4/189 (2%)	22/187 (11%)	128/13 (91%)	146/48 (75%)	217/21 (91%)
BMI, $kg \cdot m^{-2}$	27.1±4.8	23.6±4.3	27.0±5.0	27.2±4.6	26.4±4.6	25.8±5.1
Augmentation therapy: yes/no, %	5/258 (2%)	9/184 (5%)	1/208 (0%)	20/121 (14%)	62/132 (32%)	186/52 (78%)
Lung emphysema: yes/no, %	18/245 (7%)	8/185 (4%)	23/186 (11%)	98/43 (70%)	150/44 (77%)	209/29 (88%)
Neutrophils, %	53.1±11.6	54.4±9.8	52.6±9.8	60.0±12.0	56.7±9.5	59.6±11.8
Eosinophils, %	3.2±2.4	3.0±2.2	3.0±1.8	2.7±1.8	3.0±2.1	2.8±2.2
Dyspnoea: yes/no, %	76/187 (29%)	33/160 (17%)	62/147 (30%)	125/16 (89%)	168/26 (87%)	229/9 (96%)
Index case: yes/no, %	171/91 (65%)	113/80 (59%)	127/82 (61%)	130/11 (92%)	151/42 (78%)	203/35 (85%)

Data are presented as n (%) or mean±SD, unless otherwise stated. AAT: α-1 antitrypsin; y: year; FEV_1: forced expiratory volume in 1 s; % pred: percentage relative to predicted value; D_{LCO}: diffusion capacity of the lung for carbon monoxide; CAT: COPD assessment test.

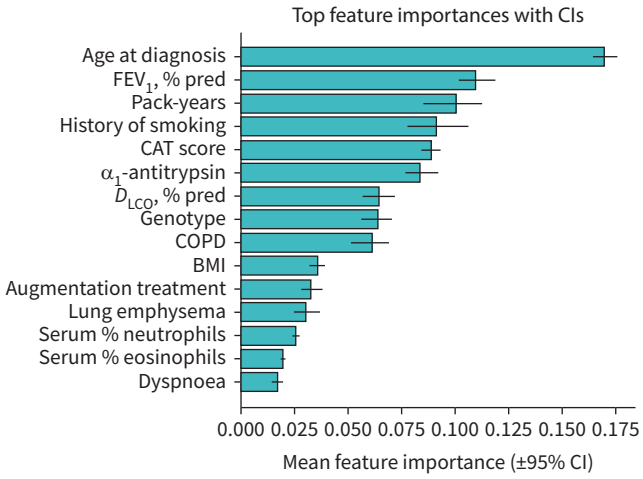

FIGURE 3 Top feature importances in differentiating the clusters. CI: confidence interval; FEV_1: forced expiratory volume in 1 s; % pred: percentage relative to predicted value; CAT: COPD assessment test; D_{LCO}: diffusion capacity of the lung for carbon monoxide; BMI: body mass index.

cluster 6, with corresponding AAT serum levels of 58.8±20.2 $mg \cdot dL^{-1}$ and 26.2±13.9 $mg \cdot dL^{-1}$, respectively. The severely reduced lung function in cluster 6 was in keeping with previous studies showing that particularly individuals with the ZZ genotype represent 95% of severe AATD with clinical manifestations due to the higher misfolding and polymerisation propensity of this AAT variant [37–40].

Moreover, smoking has already been shown to accelerate lung parenchymal damage, particularly for the ZZ genotype [17, 19–21]. This is apparent in clusters 5 and 6 that grouped individuals having

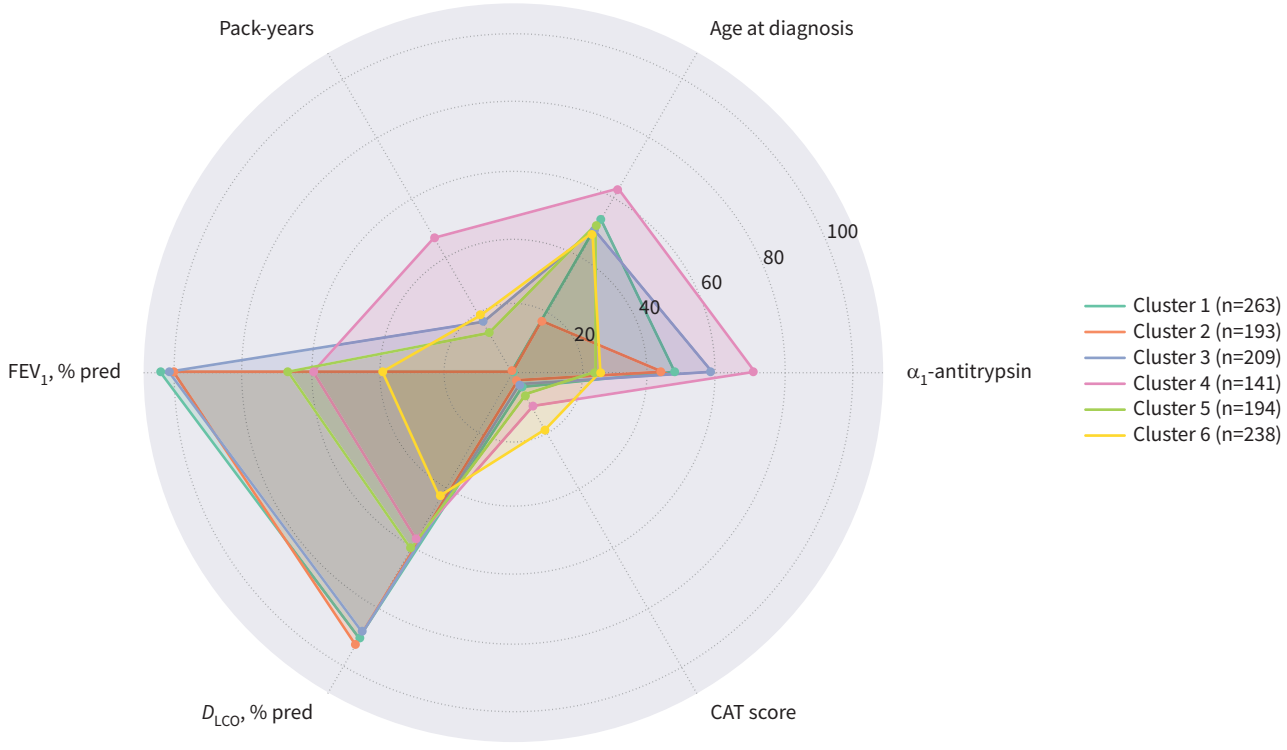

FIGURE 4 Radar plot with raw values of the six most distinguishing continuous variables for cluster differentiation. FEV_1: forced expiratory volume in 1 s; % pred: percentage relative to predicted value; D_{LCO}: diffusion capacity of the lung for carbon monoxide; CAT: COPD assessment test.

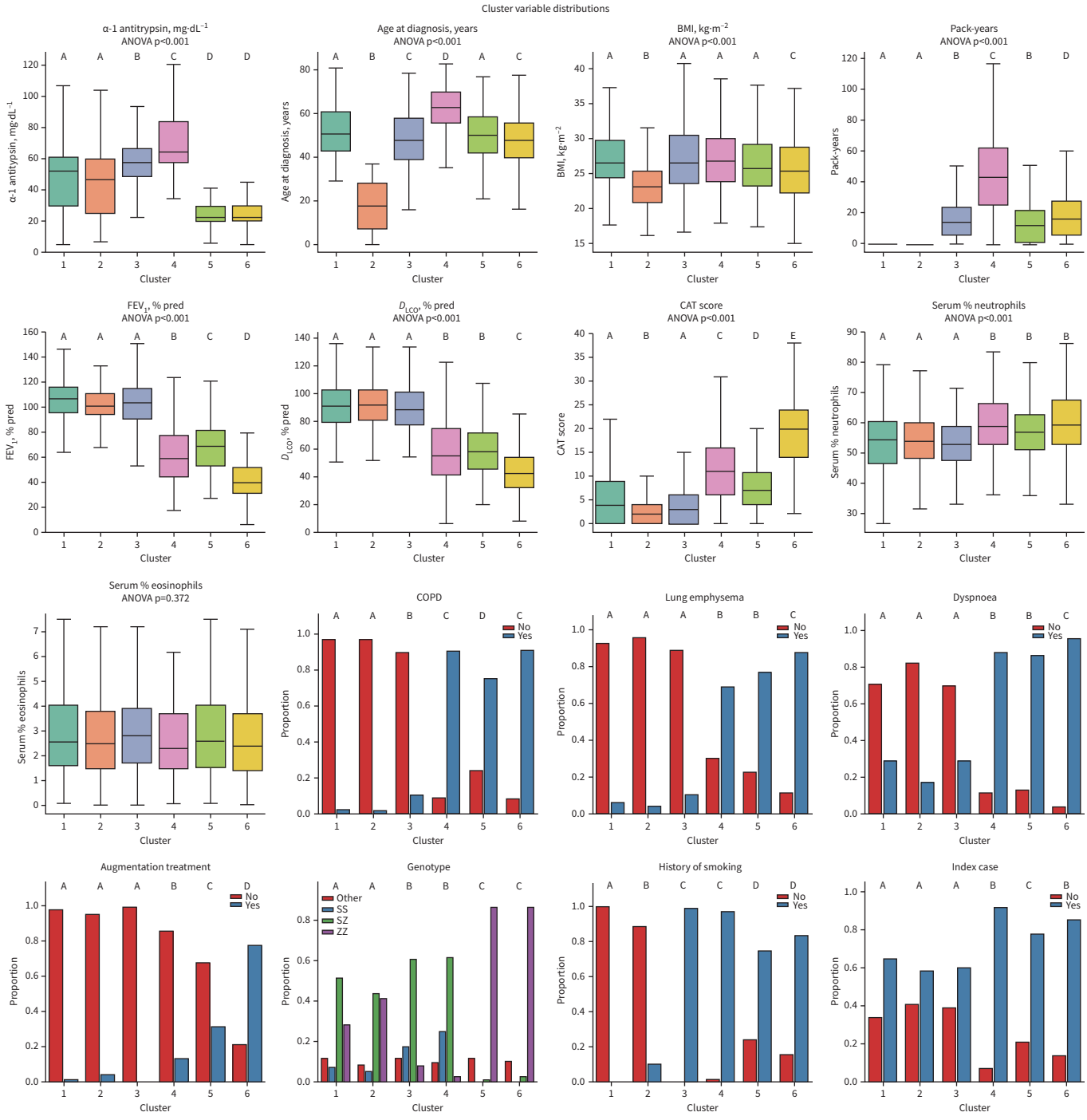

FIGURE 5 Six clusters of clinical phenotypes in α-1 antitrypsin deficiency (AATD). Statistically significant differences between clusters are indicated by a different letter (A–E). FEV_1: forced expiratory volume in 1 s; % pred: percentage relative to predicted value; D_{LCO}: diffusion capacity of the lung for carbon monoxide; CAT: COPD assessment test.

predominantly the ZZ genotype and had a more severely reduced lung function with regard to pack-years compared with clusters having predominantly the SZ genotype. Remarkably, although clusters 5 and 6 had the same prevailing proportion of individuals with the ZZ genotype, similar AAT serum levels and a similar age at AATD diagnosis, individuals in cluster 6 had a significantly lower lung function. While there are slightly more pack-years reported by individuals in cluster 6, the marked disparity in lung function appears to have a greater impact in differentiating the clusters and suggests the involvement of other unidentified contributing factors.

Altogether, we have identified six AATD clinical phenotypes using baseline data from the EARCO registry. Our findings demonstrate that besides the AATD-associated genotype and corresponding AAT serum levels, other features were even more prominent for defining clinical phenotypes, being primarily age at diagnosis, lung function and tobacco consumption. Strikingly, liver disease was not among the key distinguishing features, likely due the low inclusion number of individuals with AATD having liver disease, as the EARCO registry is mainly respiratory-driven. Furthermore, the differences and similarities between clusters suggest the presence of other, still undiscovered factors that contribute to the different clinical presentations.

A major strength of this study is the large sample size, made possible through the use of the EARCO registry. However, it should be noted that there are inherent limitations associated with multinational clinical registries such as the EARCO registry, including missing data and retrospective design. Considering that the EARCO registry is continuously expanding, future analyses should be conducted as additional data, particularly longitudinal follow-up data, become available to validate the current clinical phenotypes and to provide potential prognostic value.

Moreover, by employing the K-prototypes algorithm, we were able to include both continuous and categorical data simultaneously, as well as ensure a robust and objective differentiation of clinical phenotypes. Besides, the random forest classifier allowed to determine dominant features in differentiating the clusters. However, the study was limited to the baseline data of the EARCO study and, therefore, longitudinal trajectories of the clusters could not be assessed, warranting careful consideration of the findings. After longitudinal analysis and prospective external validation, the prediction model may serve as a tool to allocate new individuals to these clusters based on their clinical presentation. This could become particularly valuable if these clusters prove to have prognostic relevance.

Conclusion

Altogether, we identified six clusters of AATD clinical phenotypes mainly differentiated by age at AATD diagnosis, lung function and tobacco consumption, in addition to the AATD-associated genotype. While the six AATD clinical phenotypes identified in this study are currently mainly descriptive, they hold promise for providing prognostic value, especially when integrated with longitudinal data such as lung function decline, by enhancing our understanding of disease progression and allowing to assess the effectiveness of targeted therapies in the distinct phenotypes. Overall, this study underscores that, beyond genotype, a multitude of other factors significantly influence AATD clinical phenotypes. Hence it highlights the necessity of integrating (epi)genetic factors with environmental and behavioural traits, in future fundamental and translational research to unravel the complex mechanisms underlying patient heterogeneity.

Acknowledgements: The authors would like to thank the individuals with α_1-antitrypsin deficiency who participated in this study and the EARCO study investigators. We acknowledge Elise Heuvelin and Valerija Arsovski from the European Respiratory Society (Lausanne, Switzerland) for their support in the management of EARCO, and Andrea Forés and Mireia Bonet (Bioclever, Barcelona, Spain) for their support in EARCO data monitoring.

Provenance: Submitted article, peer reviewed.

EARCO study investigators: Mariano Fernandez-Acquier, Andrés L. Echazarreta (Argentina), Georg-Christian Funk, Karen Schmid-Scherzer (Austria), Wim Janssens, Silvia Pérez-Bogerd (Belgium), Kenneth R. Chapman (Canada), Leidy Prada (Colombia), Ana Hecimovic (Croatia), Eva Bartošovská (Czech Republic), Alan Altraja, Jaanus Martti (Estonia), Eric Y.E. Derom, Maeva Zysman, Jean-François Mornex, Martine Reynaud-Gaubert (France), Timm Greulich, Felix J.F. Herth, Franziska Trudzinski, Rembert Koczulla, Matthias Welsner (Germany), Gerry McElvaney (Ireland), Angelo G. Corsico, Ilaria Ferrarotti, Simone Scarlata, Mario Malerba, Luciano Corda (Italy), Jan Stolk, Emily F. van 't Wout (Netherlands), Joanna Chorostowska-Wynimko (Poland), Catarina Guimaraes, Maria Sucena, Ana Caldas, Raquel Marçoa, Isabel Ruivo dos Santos, Bebiana Conde, Maria Joana Reis Amado, Maia Da Silva, Rita Boaventura, Cristina Santos, Gabriela Santos, Filipa Costa, Teresa Martin, Joana Gomes, Sonia Isabel Silva Guerra (Portugal), Ruxandra Ulmeanu (Romania), María Torres-Duran, Marc Miravitlles, Miriam Barrecheguren, Juan Luis Rodriguez-Hermosa, Myriam Calle-Rubio, José María Hernández-Pérez, José Luis López-Campos, Francisco Casas-Maldonado, Ana Bustamante, Carlota Rodriguez-García, Marta García-Clemente, Cruz González, Eva Tabernero, Lourdes Lázaro, Virginia Almadana, Mar Fernández-Nieto, Francisco Javier Michel de la Rosa, Carlos Martínez-Rivera, Layla Diab, María Isabel Parra, Nuria Rodríguez-Lázaro, Susana Martínez, Rosanel Amaro, Ramon-Antonio Tubio (Spain), Hanan Tanash, Eeva Piitulainen (Sweden), Christian F. Clarenbach (Switzerland), Serap Argun Baris, Dilek Karadogan, Sebahat Genç, Ouksel Hakima (Turkey), Alice M. Turner, Beatriz Lara, David G. Parr, Charlotte Bolton, John Hurst, Ravi Mahadeva and Nicholas Hopkinson (UK). EARCO Steering committee:

Christian F. Clarenbach and Marc Miravitlles (co-chairs), David G. Parr, Catarina Guimaraes, Hanan Tanash, Karen O'Hara, Marion Wilkens, Alice M. Turner, Jens-Ulrik Stæhr Jensen, Maria Torres-Duran and Angelo Corsico.

Author contributions: A. Somers was responsible for critical interpretation of data and writing the original draft. K. Verstraete was responsible for formal data analysis and statistical analyses. W. Janssens and S. Everaerts were responsible for conception, design and supervision. E. Demesmaeker, M. Torres-Duran, A.M. Turner, H. Tanash, C. Rodríguez-García, J-U. Stæhr Jensen, A.G. Corsico, J.L. López-Campos, K.R. Chapman, C.F. Clarenbach, C. Guimarães, E. Bartošovská, J.M. Hernández-Pérez, J.L. Rodriguez-Hermosa, B. Lara, M. Miravitlles, W. Janssens and S. Everaerts were responsible for data acquisition within the EARCO registry. K. Verstraete, E. Demesmaeker, M. Torres-Duran, A.M. Turner, H. Tanash, C. Rodríguez-García, J-U. Stæhr Jensen, A.G. Corsico, J.L. López-Campos, K.R. Chapman, C.F. Clarenbach, C. Guimarães, E. Bartošovská, J.M. Hernández-Pérez, J.L. Rodriguez-Hermosa, B. Lara, M. Miravitlles, W. Janssens and S. Everaerts were responsible for critically reviewing and editing the manuscript. All authors have approved and contributed towards the final written manuscript.

Conflict of interest: A. Somers is supported by an FWO research grant (1123626N). M. Torres-Duran reports grants from CSL Behring and Grifols, consulting fees from CSL Behring and Grifols, and support for attending meetings from CSL Behring, Grifols, Chiesi and FAES Farma. A.M. Turner is funded by the National Institute for Health and Care Research (NIHR) Midlands Patient Safety Research Collaboration and West Midlands Applied Research Collaborative (the views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care; grants, consulting or speaker fees from AstraZeneca, Beam, AiRNA, GlaxoSmithKline, Boehringer Ingelheim, Chiesi, CSL Behring, GSK, Takeda, Vertex and Grifols Biotherapeutics; and support for attending meetings from AstraZeneca. C. Rodríguez-García reports speaker fees from AstraZeneca, GlaxoSmithKline, Grifols, Chiesi, and CSL Behring; payment for expert testimony from Chiesi; and support for attending meetings from Chiesi and Grifols. A.G. Corsico reports speaker fees and honoraria for participation on advisory board from CSL Behring, and honoraria for manuscript writing from Grifols. J.L. López-Campos reports grants from Menarini and Grifols; consultancy fees from GlaxoSmithKline, Menarini, Zambon, Gebro and Sanofi; payment or honoraria for lectures, presentations, manuscript writing or educational events from AstraZeneca, Bial, Chiesi, CSL Behring, FAES, Gebro, GlaxoSmithKline, Grifols, GSK, Menarini, Sanofi and Zambon; and support for attending meetings from Grifols, Chiesi and Gebro. C.F. Clarenbach reports consultancy fees from AstraZeneca, Boehringer Ingelheim, CSL Behring, GlaxoSmithKline, Ideogen, Novartis, Sanofi, OM Pharma, MSD, Grifols and Vifor; speaker fees from AstraZeneca, Boehringer Ingelheim, CSL Behring, GlaxoSmithKline, Novartis, Sanofi, OM Pharma, MSD, Grifols and Vifor; and support for attending meetings from Sanofi and AstraZeneca. C. Guimarães reports speaker fees from CSL Behring. J.M. Hernández-Pérez reports consulting fees from Grifols and CSL Behring; speaker fees from AstraZeneca, Bial, CSL Behring, FAES laboratory, GlaxoSmithKline and Grifols; support for attending meetings from Grifols, Bial, FAES, Gebro and CSL Behring; and honoraria for participation on advisory board from Grifols and CSL Behring. J.L. Rodriguez-Hermosa reports speaker fees from Bial, Boehringer Ingelheim, GlaxoSmithKline, Zambon, Grifols and Chiesi; support for attending meetings from Chiesi and Bial; and participated on the advisory board of Bial and CSL Behring. M. Miravitlles reports grants from Grifols; consultancy fees from AstraZeneca, Atriva Therapeutics, Boehringer Ingelheim, Chiesi, GlaxoSmithKline, Bial, Gebro Pharma, Kamada, CSL Behring, Laboratorios Esteve, Ferrer, Inhibrx, Menarini, Mereo Biopharma, Verona Pharma, TEVA, Spin Therapeutics, Specialty Therapeutics, AIRNA, GondolaBio, Palobiofarma SL, Takeda, pH Pharma, Novartis, Sanofi-Regeneron, Zambon, Zentiva and Grifols; speaker fees from AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Kamada, Chiesi, Cipla, Menarini, Rovi, Bial, Sandoz, Takeda, Zambon, CSL Behring, Specialty Therapeutics, Tabuk Pharmaceuticals, Glenmark Pharmaceuticals, Grifols, Sanofi-Regeneron and Novartis; support for attending meetings from Boehringer Ingelheim, Menarini, Chiesi, Bial, CSL Behring and Grifols; and participation on a data safety monitoring board for Mereo. W. Janssens reports consultancy fees from GlaxoSmithKline, AstraZeneca, Sanofi, Roche and Chiesi; payment or honoraria for lectures, presentations, manuscript writing or educational events from GlaxoSmithKline, AstraZeneca, Sanofi and Chiesi; and research grants from GlaxoSmithKline, Chiesi and AstraZeneca. S. Everaerts reports consultancy fees from GlaxoSmithKline; payment or honoraria for lectures, presentations, manuscript writing or educational events from GlaxoSmithKline, PulmonX, Chiesi and AstraZeneca; support for attending meetings from GlaxoSmithKline, PulmonX, Sanofi and AstraZeneca; and a research grant from Chiesi. The remaining authors report no conflicts of interest.

Support statement: The International EARCO registry is funded by unrestricted grants from Grifols, CSL Behring, Kamada, pH Pharma, Sanofi and Takeda to the European Respiratory Society.

References

1 Strnad P, McElvaney NG, Lomas DA. Alpha1-antitrypsin deficiency. *N Engl J Med* 2020; 382: 1443–1455.
2 Janciauskiene SM, Bals R, Koczulla R, *et al.* The discovery of α1-antitrypsin and its role in health and disease. *Respir Med* 2011; 105: 1129–1139.

3 Laurell CB, Eriksson S. The electrophoretic alpha1-globulin pattern of serum in alpha1-antitrypsin deficiency. *Scand J Clin Lab Invest* 1963; 15: 132–140.
4 Adams TS, Schupp JC, Poli S, *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1983
5 Basil MC, Cardenas-Diaz FL, Kathiriya JJ, *et al.* Human distal airways contain a multipotent secretory cell that can regenerate alveoli. *Nature* 2022; 604: 120–126.
6 Habermann AC, Gutierrez AJ, Bui LT, *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1972
7 Vieira Braga FA, Kar G, Berg M, *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat Med* 2019; 25: 1153–1163.
8 Reyfman PA, Walter JM, Joshi N, *et al.* Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am J Respir Crit Care Med* 2019; 199: 1517–1536.
9 Abo KM, Merritt C, Basil MC, *et al.* Pulmonary cellular toxicity in alpha-1 antitrypsin deficiency. *Chest* 2024; 166: 472–479.
10 Janciauskiene S. Conformational properties of serine proteinase inhibitors (serpins) confer multiple pathophysiological roles. *Biochim Biophys Acta* 2001; 1535: 221–235.
11 Jonigk D, Al-Omari M, Maegel L, *et al.* Anti-inflammatory and immunomodulatory properties of α1-antitrypsin without inhibition of elastase. *Proc Natl Acad Sci USA* 2013; 110: 15007–15012.
12 Campbell EJ, Campbell MA, Boukedes SS, *et al.* Quantum proteolysis by neutrophils: Implications for pulmonary emphysema in α1-antitrypsin deficiency. *J Clin Invest* 1999; 104: 337–344.
13 Senior RM, Tegner H, Kuhn C, *et al.* The induction of pulmonary emphysema with human leukocyte elastase. *Am Rev Respir Dis* 1977; 116: 469–475.
14 Guyot N, Wartelle J, Malleret L, *et al.* Unopposed cathepsin G, neutrophil elastase, and proteinase 3 cause severe lung damage and emphysema. *Am J Pathol* 2014; 184: 2197–2210.
15 Rabe KF, Watz H. Chronic obstructive pulmonary disease. *Lancet* 2017; 389: 1931–1940.
16 Mahadeva R, Atkinson C, Li Z, *et al.* Polymers of Z α1-antitrypsin co-localize with neutrophils in emphysematous alveoli and are chemotactic in vivo. *Am J Pathol* 2005; 166: 377–386.
17 Mulgrew AT, Taggart CC, Lawless MW, *et al.* Z α1-antitrypsin polymerizes in the lung and acts as a neutrophil chemoattractant. *Chest* 2004; 125: 1952–1957.
18 Lomas DA, Evans DL, Finch JT, *et al.* The mechanism of Z α1-antitrypsin accumulation in the liver. *Nature* 1992; 357: 605–607.
19 Carrell RW, Jeppsson JO, Laurell CB, *et al.* Structure and variation of human α1–antitrypsin. *Nature* 1982; 298: 329–334.
20 Brantly ML, Paul LD, Miller BH, *et al.* Clinical features and history of the destructive lung disease associated with alpha-1-antitrypsin deficiency of adults with pulmonary symptoms. *Am Rev Respir Dis* 1988; 138: 327–336.
21 Alam S, Li Z, Janciauskiene S, *et al.* Oxidation of Z α1-antitrypsin by cigarette smoke induces polymerization. *Am J Respir Cell Mol Biol* 2011; 45: 261–269.
22 Sinden NJ, Baker MJ, Smith DJ, *et al.* α-1-Antitrypsin variants and the proteinase/antiproteinase imbalance in chronic obstructive pulmonary disease. *Am J Physiol Lung Cell Mol Physiol* 2015; 308: L179–L190.
23 Gooptu B, Dickens JA, Lomas DA. The molecular and cellular pathology of α1-antitrypsin deficiency. *Trends Mol Med* 2014; 20: 116–127.
24 de Serres FJ, Blanco I. Prevalence of α1-antitrypsin deficiency alleles PI*S and PI*Z worldwide and effective screening for each of the five phenotypic classes PI*MS, PI*MZ, PI*SS, PI*SZ, and PI*ZZ: a comprehensive review. *Ther Adv Respir Dis* 2012; 6: 277–295.
25 American Thoracic Society, European Respiration Society. American Thoracic Society/European Respiratory Society statement: Standards for the diagnosis and management of individuals with alpha-1 antitrypsin deficiency. *Am J Respir Crit Care Med* 2003; 168: 818–900.
26 Miravitlles M, Dirksen A, Ferrarotti I, *et al.* European Respiratory Society statement: diagnosis and treatment of pulmonary disease in α1-antitrypsin deficiency. *Eur Respir J* 2017; 50: 1700610.
27 Zorzetto M, Russi E, Senn O, *et al.* SERPINA1 gene variants in individuals from the general population with reduced α1-antitrypsin concentrations. *Clin Chem* 2008; 54: 1331–1338.
28 Aktas A, Walsh D, Rybicki L. Symptom clusters: Myth or reality? *Palliat Med* 2010; 24: 373–385.
29 Haldar P, Pavord ID, Shaw DE, *et al.* Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008; 178: 218–224.
30 Greulich T, Altraja A, Barrecheguren M, *et al.* Protocol for the EARCO Registry: a pan-European observational study in patients with α1-antitrypsin deficiency. *ERJ Open Res* 2020; 6: 00181–2019.
31 Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl Discov* 1998; 2: 283–304.
32 Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
33 Van Rijsbergen CJ. Information retrieval, 2nd ed. London, Butterworth-Heinemann, 1979.

34 Tanash HA, Ekström M, Rönmark E, *et al.* Survival in individuals with severe alpha 1-antitrypsin deficiency (PiZZ) in comparison to a general population with known smoking habits. *Eur Respir J* 2017; 50: 1700198.

35 Tanash HA, Nilsson PM, Nilsson JA, *et al.* Clinical course and prognosis of never-smokers with severe alpha-1-antitrypsin deficiency (PiZZ). *Thorax* 2008; 63: 1091–1095.

36 Mostafavi B, Piitulainen E, Tanash HA. Survival in the Swedish cohort with alpha-1-antitrypsin deficiency, up to the age of 43–45 years. *Int J Chron Obstruct Pulmon Dis* 2019; 14: 525–530.

37 Blanco I, Diego I, Castañón C, *et al.* Estimated worldwide prevalence of the PI*ZZ alpha-1 antitrypsin genotype in subjects with chronic obstructive pulmonary disease. *Arch Bronconeumol* 2023; 59: 427–434.

38 Gooptu B, Lomas DA. Polymers and inflammation: disease mechanisms of the serpinopathies. *J Exp Med* 2008; 205: 1529–1534.

39 Lomas DA, Mahadeva R. α1-Antitrypsin polymerization and the serpinopathies: pathobiology and prospects for therapy. *J Clin Invest* 2002; 110: 1585–1590.

40 Dafforn TR, Mahadeva R, Elliott PR, *et al.* A kinetic mechanism for the polymerization of α1-antitrypsin. *J Biol Chem* 1999 Apr; 274: 9548–9555.