Machine learning-derived asthma phenotypes in a representative Swedish adult population

Muwada Bashir Awad Bashir[1], Daniil Lisik[1], Saliha Selin Özuygur Ermis[1], Rani Basna[2], Reshed Abohalaka[1], Selin Ercan[1], Helena Backman[3], Teet Pullerits[4], Roxana Mincheva[1], Göran Wennergren[5], Madeleine Rådinger[1], Jan Lötvall[1], Linda Ekerljung[1], Hannu Kankaanranta[1], and Bright Nwaru[1]

[1]University of Gothenburg
[2]Lunds universitet Institutionen for kliniska vetenskaper Malmo
[3]Umea universitet Institutionen for folkhalsa och klinisk medicin
[4]Goteborgs universitet Avdelningen for invartesmedicin och klinisk nutrition
[5]Goteborgs universitet Pediatrik

April 19, 2024

Abstract

Background Asthma is a heterogenous airway disease characterized by multiple phenotypes. Unbiased identification of these phenotypes is paramount for optimizing asthma management. **Objectives** To identify and characterize asthma phenotypes based on a broad set of attributes using a novel machine learning approach in a representative sample of Swedish adults. **Methods** Deep learning clustering was used to derive asthma phenotypes in a sample of 1,895 subjects aged 16-75, drawn from the ongoing West Sweden Asthma Study. The algorithm integrated 47 variables encompassing demographics, risk factors, asthma triggers, pulmonary function, disease severity, allergy, and comorbidity profiles. The optimal clustering solution was selected by combining statistical metrics and clinical interpretation. **Results** A four-cluster solution was determined to reliably represent the data, resulting in distinct phenotypes described as: (1) troublesome, late-onset, non-atopic asthma with smoking (n=458, 24.2%); 2) female-dominated early adult-onset asthma (n=545, 28.7%); 3) adult-onset asthma with high inflammation (n=358, 18.9%); and 4) early-onset, mild, atopic asthma (n=534, 28.2%). The phenotypes also differed with respect to demographics, risk factors, asthma triggers, pulmonary function, symptom profiles, and markers of inflammation. Current asthma was more common in phenotypes with later age of asthma onset than phenotypes with early onset. **Conclusion** Four clinically meaningful asthma phenotypes, distinguishable by age of onset, severity, risk factors, and prognosis, were found in Swedish adults. This provides a setting for future research to profile the immunological basis of the phenotypes, and further our understanding of their pathophysiology, therapeutic possibilities, future clinical outcomes, and societal burden.

Machine learning-derived asthma phenotypes in a representative Swedish adult population

Short title: Machine learning-derived asthma phenotypes

Muwada Bashir Awad Bashir,[1] Daniil Lisik,[1] Saliha Selin Ozuygur Ermis,[1]Rani Basna[2], Reshed Abohalaka,[1]Selin Ercan,[1] Helena Backman,[3] Teet Pullerits,[4] Roxana Mincheva,[1] Göran Wennergren,[5] Madeleine Rådiinger,[1]Jan lotvall,[1] Linda Ekerljung,[1]Hannu Kankaanranta,[1,6,7] Bright I. Nwaru[1,8].

[1]Krefting Research Centre, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[2]Department of Clinical Sciences in Malmö, Division of Geriatric Medicine, Lund University, Sweden

[3] Department of Public Health and Clinical Medicine, Section of Sustainable Health/the OLIN Unit, Umeå University, Umeå, Sweden.

[4]Krefting Research Centre, University of Gothenburg, Gothenburg, Sweden; Respiratory Medicine and Allergy, Department of Internal Medicine & Clinical Nutrition, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden; Section of Allergology, Sahlgrenska University Hospital, Gothenburg, Sweden.

[5]Department of Pediatrics, University of Gothenburg, Gothenburg, Sweden.

[6]Department of Respiratory Medicine, Seinäjoki Central Hospital, Seinäjoki, Finland.

[7]Tampere University Respiratory Research Group, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

[8]Wallenberg Centre for Molecular and Translational Medicine, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden.

Word count: 3376

Conflict of interest

HK reports personal fees for lectures and consulting from AstraZeneca, Boehringer-Ingelheim, Chiesi, COVIS Pharma, GSK, Medscape, MSD, Novartis, Orion Pharma and Sanofi. All other authors of this work declare no conflict of interest related to current work.

Funding

VBG Group Herman Krefting Foundation for Asthma and Allergy Research, the Swedish Research Council, the Research Foundation of the Swedish Asthma and Allergy Association, the Nordic Epilung under the support of Nordforsk, the Swedish Heart-Lung Foundation, and the Swedish government under the ALF agreement between the Swedish government and the county councils (Västra Götaland). None of the sponsors had any involvement in the planning, execution, drafting or write-up of this study.

Authors' contribution

BN conceived the idea of the study. MB processed the data, performed the analysis, drafted the manuscript, and designed the figures under supervision of BN. LE, GW, HK, RB contributed to the implementation of the data processing and analysis. BN contributed to the design and implementation of the research and data collection. All authors discussed the results and contributed to the final manuscript. This study was supervised by BN and HK.

Ethical approval

The study was approved by the regional ethical review board in Gothenburg, Sweden. Ethical approval number Dnr: 052-16.

Abstract

Background

Asthma is a heterogenous airway disease characterized by multiple phenotypes. Unbiased identification of these phenotypes is paramount for optimizing asthma management.

Objectives

To identify and characterize asthma phenotypes based on a broad set of attributes using a novel machine learning approach in a representative sample of Swedish adults.

Methods

Deep learning clustering was used to derive asthma phenotypes in a sample of 1,895 subjects aged 16-75, drawn from the ongoing West Sweden Asthma Study. The algorithm integrated 47 variables encompassing demographics, risk factors, asthma triggers, pulmonary function, disease severity, allergy, and comorbidity profiles. The optimal clustering solution was selected by combining statistical metrics and clinical interpretation.

Results

A four-cluster solution was determined to reliably represent the data, resulting in distinct phenotypes described as: (1) troublesome, late-onset, non-atopic asthma with smoking (n =458, 24.2%); 2) female-dominated early adult-onset asthma (n =545, 28.7%); 3) adult-onset asthma with high inflammation (n =358, 18.9%); and 4) early-onset, mild, atopic asthma (n =534, 28.2%). The phenotypes also differed with respect to demographics, risk factors, asthma triggers, pulmonary function, symptom profiles, and markers of inflammation. Current asthma was more common in phenotypes with later age of asthma onset than phenotypes with early onset.

Conclusion

Four clinically meaningful asthma phenotypes, distinguishable by age of onset, severity, risk factors, and prognosis, were found in Swedish adults. This provides a setting for future research to profile the immunological basis of the phenotypes, and further our understanding of their pathophysiology, therapeutic possibilities, future clinical outcomes, and societal burden.

Keywords: asthma, clinical, deep learning, phenotypes, population

introduction

Asthma is a heterogeneous disease, and a clear-cut characterization of its various phenotypes has historically remained daunting for both clinical practice and research purposes [1-3]. Conventionally, asthma phenotypes have been defined based on timing of onset, atopic origin, eosinophilic inflammation, and presence of obesity, to name a few [4]. Such phenotypic characterization has been described as primarily based on clinical insights and experiences of the attending clinician. However, it has been suggested that such asthma phenotyping is largely subjective as the classification may vary from clinician to clinician[4,5]. Additionally, asthma phenotyping has mostly been attempted in selected cohorts, example hospital-based asthma patients or those with severe asthma, with less data from population-representative samples.

The advancements being made by computational science at elucidating biological processes have been welcomed in the field of asthma, particularly in identifying asthma phenotypes [4-6]. In this context, various features of asthma are inputted into algorithms that learn from unlabelled data, with less artefact bias, to produce meaningful asthma phenotypes. This data-driven approach is believed to be more objective and can, with relevant clinical inputs, produce phenotypes that are clinically meaningful [4,5,7]. Characterizing asthma at a more granular level is in parallel with efforts towards precision medicine, subsequently enabling prevention and optimal, tailored management [8].

In this work, by including a broad range of clinical, biological, and epidemiological parameters that are relevant to asthma, we employed a novel machine learning approach to identify and describe asthma phenotypes

in an adult representative sample in western Sweden.

Methods

Study population

The study sample was derived from the ongoing West Sweden Asthma Study (WSAS), a longitudinal cohort study investigating different aspects of airway diseases among a representative sample of adults of western Sweden. WSAS started in 2008 in which a random sample of 30,000 individuals aged 16-75 years were invited to participate in a postal survey. Of those who responded to the questionnaire, 2,000 randomly selected subjects and 1,524 individuals with self-reported asthma were further invited to undergo detailed clinical investigations, conducted between 2009 and 2012 (WSAS I). A total of 2,006 subjects participated in this initial clinical examination. In 2016, a non-overlapping sample from western Sweden was added to WSAS, using the same survey and clinical investigation methods. Put together, a total of 3,101 individuals underwent clinical investigations before the start of COVID-19, at which time clinical data collection was put on hold due to the pandemic, of which 1,895 subjects who had ever had asthma were included in the current phenotyping work. A detailed description of the WSAS cohort design and characteristics has been reported previously[9].

Assessment and measurements

Definition of asthma

Asthma was defined as self-reported history of ever having asthma or physician-diagnosed asthma. This definition was used to capture the aspect of timing of asthma onset. Thus, a follow-up question "at what age did your asthma start" was included.

Variables included in deriving the phenotypes

We included a list of 47 variables, which were selected based on clinical experience and previous studies. This comprehensive list of variables was included to capture a full representation of the clinical presentation, pathophysiological mechanisms, and potential risk factors related to asthma. The variables were grouped as follows: (1) demographics/triggers/risk factors; (2) symptom profiles; (3) measures of pulmonary function; (4) disease prognosis and severity indicators; and (5) measures of inflammation. All variables were self -reported by participants except for lung function, reversibility, measures of lung diffusion capacity, and inflammation biomarkers which were measured objectively in the clinic. A detailed list of all variables is presented in the **Supplementary file** .

Statistical analysis

Handling of missing data

Missing items in the data set was imputed using multiple imputation with random forest. Random forests is an ensemble learning method, primarily used for classification and regression, which operate by constructing a multitude of decision trees at training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees [10]. When applied to data imputation, random forests leverage their inherent ability to handle non-linear relationships and interactions between variables to predict missing values with high accuracy [10,11]. The imputation process was implemented using miceRanger package in R[12]. Details regarding the imputation process can be found in the **supplementary material** .

Unsupervised clustering

The derivation of the asthma phenotypes was done using Deep Embedded Clustering (DEC) [13]. DEC is a novel approach that combines deep learning, which is an advanced form of machine learning technology, with clustering, allowing for the discovery of complex patterns in data and providing a robust, scalable solution for clustering large datasets without the need for labelled data[13]. This makes it particularly valuable for applications where the true cluster structure is unknown or hard to define a priori [13]. DEC has an advantage over traditional clustering methods because of its ability to learn a lower-dimensional representation (feature space) of data using deep autoencoders. This feature space is more suitable for clustering due to compact representation at lower dimensionality, allowing DEC to outperform traditional methods that either do not involve feature learning or rely on simpler, linear dimensionality reduction techniques[13]. Secondly, DEC's iterative optimization process that utilizes distance metrices to optimize both the feature representation and cluster assignments in a way that traditional methods, such as k-means or spectral clustering, cannot[13]. These qualities make DEC particularly effective for complex datasets, offering improved clustering accuracy, efficiency in handling large datasets.

After the data was processes, The R package NbClust was used to decide the optimal number of clusters using voting consensus methods[14]. Additionally, the optimal number of cluster was confirmed using Monte Carlo reference based consensus clustering approach [15], implemented through M3C R package[16]. The output was further fed into the DEC algorithm to perform the clustering. The cluster were later validated using prediction strength approach. The final numbers proposed by such metrics were then evaluated in conjunction with clinical experience before a final determination of the optimal number of clusters were decided to represent the data. The cluster solution determined were then named based on their distribution with regards to the variables used to derive the clusters. A detailed statistical implementation is presented in the**Supplementary file** .

Other statistical analysis

Continuous data were expressed as means and standard deviations (SD). Group comparisons were performed by 1-way analysis of variance with the Tukey post hoc test, the Kruskal- Wallis test, or the chi-square, for categorical and continuous variables as suitable. Graphical representation of variation between cluster was performed through radar plots, where categorical variables were represented as proportions.

Reproducibility and data availability

The full information describing the analysis approach implemented as well as the analysis codes are available in the online repository (https://github.com/ranibasna/AsthmaClustering).

Results

Baseline characteristics of asthma versus non-asthma subjects in the sample

Of the 3,101 subjects who participated in the WSAS clinical investigations prior to COVID-19, 1,206 had never had asthma, while the remaining 1,895 who had ever had asthma were included in the phenotype derivation. General characteristics of subjects with asthma and without asthma, respectively, can be seen in Table 1. Those who had ever had asthma were younger, had a higher proportion of women, were more obese, had poorer lung function, more inflammation, and more respiratory symptoms, and were less atopic, had a higher proportion of family history of asthma/allergy, less rhinitis, and more COPD than those who had never had asthma.

Derived asthma phenotypes and their descriptive names

Based on the DEC computational metrics, a three-cluster solution had the best metrics, but when subjected to a clinical review, we determined that a four-cluster solution was the most meaningful and valid representation of the data, both clinically and scientifically, capturing the diversity of phenotypes, hence we chose the four-cluster solution. The clusters were given the following descriptive names:

- Phenotype 1 (n=458, 24.2%): troublesome late-onset, non-atopic asthma with more women and smokers
- Phenotype 2 (n=545, 28.7%): Female-dominated early adult-onset asthma
- Phenotype 3 (n=358, 18.9%): Adult-onset asthma with high inflammation
- Phenotype 4 (n=534, 28.2%): Early-onset, mild, atopic asthma

Detailed information on differences between clusters can be found in the supplementary material.

Demographic, risk factors and asthma triggers in derived phenotypes

Cluster 1 had the highest average age at asthma onset 35, while Cluster 4 had the youngest average age at asthma onset 13 **(Table 1).**Cluster 1 had the highest average calendar age (65 and the highest average BMI (28.5) than the other clusters. While it had a lower urbanization rate (38.9%) than the other clusters, the proportion of smokers and average pack-year history was the highest. Cluster 1 also had the highest proportion of individuals exposed to smoking at home and at work than other clusters (**Table 1** and **Figure 1:A and B**). On the other hand, Cluster 2 had the second youngest average age at asthma onset, was the most female-dominated cluster, the second least urbanized, highest proportion of never smokers. In this cluster, the subjects had experienced more household mould, plastic-carpet or water damage, the highest rate of exertion-induced respiratory symptoms, and the highest rate of infection-triggered asthma symptoms than other clusters (**Figure 1: A and B**). Cluster 3 had the second oldest age at asthma onset and had a higher proportion of males than other clusters (**Figure 1** and **Figure 2**). Cluster 4 had the youngest average age at asthma onset, had the lowest average calendar age and BMI, more equally distributed between men and women, most urbanized, lowest pack-years of smoking, lowest rates of smoking exposure at home and at work (29.8%), and the lowest rates of exercise and infection as asthma triggers (**Table 1** and **Figure 1: A**).

Derived asthma phenotypes by markers of inflammation

Cluster 3 had the highest FeNO levels and eosinophil count. Cluster 1 had the highest neutrophil count. Cluster 4 had the lowest FeNO levels and neutrophil count. There were no marked differences between cluster 1, 2 and 4 regarding FeNO levels. (**Figure 2)** .

Derived asthma phenotypes by symptom profiles, allergic status, and comorbidities

Cluster 1 had the highest reports of respiratory symptoms but had the lowest rates of allergic sensitization and family history of asthma or allergy. It also had the highest proportion of individuals with asthma related nasal polyps, rhinitis, and COPD (**Figure 1: C**). On the other hand, cluster 4 had the least respiratory symptoms, the highest rate of allergic sensitization, and the lowest proportion of individuals with co-existing COPD (**Figure 1: C**). Clusters 2 and 3 were in-between clusters 1 and 4 in these aspects.

Derived asthma phenotypes by lung function

Cluster 1 had the lowest post-bronchodilator FEV_1/FVC ratio and the second highest reversibility after bronchodilator therapy among the clusters. Conversely, cluster 4 had the highest FEV_1/FVC ratio and the lowest reversibility score. Clusters 2 and 3 were in-between clusters 1 and 4 in lung function (**Figure 3**).

Derived asthma phenotypes by disease prognosis and severity

Cluster 1 had the highest proportion of individuals with uncontrolled asthma based on GINA classification and the highest proportion of individuals on GINA treatment steps 4 and 5. Cluster 1 also had the second-highest proportion of individuals who only used short-acting beta-agonists (SABA) or never had used asthma medication. On the other hand, cluster 4 had the lowest proportion of individuals with uncontrolled asthma, the lowest proportion of individuals on GINA treatment steps 4 and 5, and the highest proportion of individuals who only used short-acting beta-agonists (SABA) or never had been prescribed asthma medication. Clusters 2 and 3 were in-between clusters 1 and 4 in regarding asthma severity (**Figure 1: B**).

Description of derived asthma phenotypes by current asthma

The prevalence of current asthma among both men and women was higher for clusters with later onset-age of asthma than for those with early onset. For instance, while almost all individuals (both in men and women) in cluster 1 (the cluster with the highest age at asthma onset) had current asthma, the proportion was less than 80% in cluster 4 (the cluster with the earliest age of asthma onset) **(Figure 3).**

Discussion

By including a large sample of asthma subjects with ever reported or physician-diagnosed asthma and a comprehensive set of parameters, covering demographic/risk factors/triggers, clinical, and pathophysiological aspects into a novel machine learning algorithm, we could derive a four-cluster solution that captured clinically meaningful asthma phenotypes. The derived asthma phenotypes could be distinguished based on age at asthma onset, ranging from those with onset in childhood to those with onset in adulthood. They could also be distinguished on the basis of level of severity, ranging from the childhood atopic mild asthma to the late adulthood more troublesome asthma. The phenotypes could also be differentiated on the basis of several demographic/risk factors/triggers, clinical aspects, symptom profiles, and various measures of inflammation.

WSAS is representative of the adult population of western Sweden; as such our findings have a reliable generalizability to the underlying target population. The population-based sampling also constitutes an advantage over some previous studies that have relied primarily on hospital-based setting in their phenotyping [7,17-19] . With a population-based sample, the whole spectrum of asthma severity can be captured. We selected a comprehensive set of variables for the phenotyping exercise, which ensured that multiple dimensions of asthma were captured, providing an advantage over approaches that utilize fewer sets of variables or that focus primarily on clinical variables. We employed a novel and robust machine learning approach, deep embedded learning, which is particularly adept at managing complex, multi-dimensional data, offering an advantage over conventional clustering approaches. The selection and description of the derived phenotypes represent a hybrid of data science and clinical experience, ensuring that the phenotypes accurately align to both clinical and statistical expectations. Our study, however, may be limited by the absence of certain inflammatory markers, like sputum measures, which are valuable in defining asthma endotypes. In addition, the lack of certain co-morbidities could also mean potential aspects of phenotype characterization were not captured. Nevertheless, the alignment of our findings to previous studies indicates that our approaches were largely valid and reliable.

The first phenotype in our work (Cluster 1) was characterized by substantially older age, late onset and troublesome asthma with increased smoking, which had high symptom and health care use burden, compared to other clusters. It also has a higher proportion of patients that can be classified as having severe asthma based on medication usage. This phenotype overlaps with findings from previous studies among adults [20-25]. For instance, a similar phenotype derived by Kaneko et al.[22] carried same characteristics as our first phenotype. Kim et al.[23]also reported a phenotype with high airway obstruction, non-atopy, and older age. The phenotype derived by Loureiro et al. [26]was similar to ours by being late onset and severe, uncontrolled asthma, dominated by obese women, and had high eosinophil, neutrophil and monocyte counts. Different characteristics of this phenotype have been described in a similar phenotype derived by other studies, including systemic inflammation [27]. late-onset and severe asthma[28], high comorbidity burden[29], need for more medication [30], [31], and increased cigarette smoking[30],[32] .

Our second phenotype (Cluster 2) that was characterized by female dominance and early adult-onset asthma with high breathlessness and moderate symptoms, nearly normal lung function, and moderate healthcare use also closely aligns with findings from previous studies[22,26,29,33-35]. This phenotype closely mirrors the phenotype identified by Dudchenko et al.[36], which was notably sensitive to weather as a trigger, while our phenotype had exercise and infections as important triggers. Ilmarinen et al.[32] reported a similar phenotype that was described as 'female asthma' and marked by near normal lung function but being moderately symptomatic and using health care services. A similar phenotype was described by Kim et al.[23] as early adulthood-onset, mild, female asthma, featuring persistent normal lung function and a gentle disease

progression in young women.

The phenotype of adult-onset asthma with high inflammation (Cluster 3) also aligns with a phenotype found in previous studies[20,28,33,37-39]. For example, Bochnek and colleagues[20] reported a moderate asthma phenotype with elevated eosinophil levels. However, their study did not address the age of asthma onset. Boudir and colleagues also identified a moderate asthma phenotype characterized by significant bronchodilator reversibility and pronounced respiratory symptoms with a high rate of atopy, consistent with our results. Hsaio and colleagues [33] also described a similar phenotype to our findings, which was further distinguished by a history of smoking.

Similarly, our fourth phenotype of early-onset, mild asthma with atopy (Cluster 4) was frequently reported in previous studies[22,23,32,36,37,39-42]. In addition to overlapping characteristics of mild disease course, good control status, high atopy, and relatively early onset, Dudchenko et al.[36]reported high impairment on physical activity that additionally characterize this phenotype. Two studies additionally reported younger age of subjects belonging to this phenotype, which is in line with our observation of young mean age among members of this phenotype. Loza et al.[39] additionally reported this phenotype to be associated with low inflammation of high T2 cell pattern.

The first phenotype (Cluster 1), characterized by late onset troublesome asthma, with older age, high rate of smoking, COPD as a comorbidity, and reduced diffusion capacity, may point to presence of emphysematous changes. Clinically, this phenotype may present asthma and COPD co-existing in the same patient [32,43-45]. Additionally, compared to the other phenotypes, with high BMI, high proportion of females, more healthcare use, hospital emergencies, and systemic inflammation, this phenotype could also be reflecting the group of severe female obesity-related asthma with mixed inflammation patterns that have been reportedly associated with severe presentation at late age [26,32,33]. The presentation of more comorbidities amongst such a group of asthma patients also aligns with the greatest impairment to quality of life that had been associated with such phenotype previously [26].

The second phenotype (Cluster 2) that constituted a group of women with moderate asthma seemed to have better overall health because they had fewer other health problems, had low smoking rates, and generally demonstrated good lung function. This group also showed relatively moderate asthma symptoms, which might be because women tend to notice their symptoms more and seek medical help sooner [46], which is demonstrated by high utilization of emergency service among this group compared to others. Further, low smoking history, low BMI, low count of comorbidities may have influenced the good overall prognosis of this female cluster. Additionally, these women were particularly good at noticing what triggered their asthma, like changes in weather or infections, which may help them avoid these triggers and have fewer symptoms.

The asthma phenotype that typically begins in early adulthood (Cluster 3) was characterized by significant inflammation and moderate symptom severity, with a prominent feature being an elevated FeNO and eosinophil count. These are associated with type 2 immune response[39,47]. Such group with high eosinophiles and FeNo levels may represent a sensitive treatment group with ICS therapy, however they may be undertreated. Additionally, this phenotype tends to have greater exposure to smoking, which has been linked to increased eosinophilic inflammation [48]. This phenotype also exhibited a high occurrence of rhinitis and allergic conditions, such as chronic nasal problems accompanying asthma. Nonetheless, the observation that this group reported the fewest symptoms of drug-induced asthma contradicts this hypothesis.[26]

The early onset, mild atopic asthma phenotype (Cluster 4) possibly represents the traditional childhood-onset asthma characterized by high allergic sensitization, better asthma control, and low symptom burden. Childhood onset-asthma has greater propensity for remission than asthma starting in adulthood [49]. Our data suggests that this phenotype also may have the highest rate of remission as it had the lowest proportion of those who have current asthma as defined by recent symptoms and medication use, indicating that although members of this phenotype developed asthma during childhood, some of them might be transient in part of patients in this cluster.

Conclusion

In a representative sample of adults who had ever had asthma, four clinically relevant asthma phenotypes can be seen, which could be distinguished on the basis of age at asthma onset, level of severity, and other characterization, including demographics, risk factors, triggers, lung function, respiratory symptom profiles, and inflammation markers. The derived asthma phenotypes provide a novel setting for catalyzing important asthma research that will include detailed profiling of the immunological aspects of the phenotypes, modelling putative risk factors, characterizing their comorbidity profiles, and assessing future clinical outcomes and societal burden of each asthma phenotype.

References

1. Moore WC, Meyers DA, Wenzel SE, et al. Identification of Asthma Phenotypes Using Cluster Analysis in the Severe Asthma Research Program.*American journal of respiratory and critical care medicine* . 2010 2010;181(4):315-323.

2. Haldar P, Pavord ID, Shaw DE, et al. Cluster Analysis and Clinical Asthma Phenotypes. *American journal of respiratory and critical care medicine* . 2008;178(3):218-224. doi:10.1164/rccm.200711-1754oc

3. Wu W, Bleecker E, Moore W, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data.*Journal of Allergy and Clinical Immunology* . 2014 2014;133(5):1280-1288.

4. Weatherall M, Travers J, Shirtcliffe P, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis.*European Respiratory Journal* . 2009;34(4):812-818.

5. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques.*American journal of respiratory and critical care medicine* . 2013;188(11):1303-1312.

6. Khanam UA, Gao Z, Adamko D, et al. A scoping review of asthma and machine learning. *Journal of Asthma* . 2023;60(2):213-226. doi:10.1080/02770903.2022.2043364

7. Deliu M, Yavuz TS, Sperrin M, et al. Features of asthma which provide meaningful insights for understanding the disease heterogeneity.*Clin Exp Allergy* . Jan 2018;48(1):39-47. doi:10.1111/cea.13014

8. Global initative for asthma. Global Strategy For Asthma Management And Prevention, 2022. 1/6/2023, 2023. https://ginasthma.org/

9. Nwaru BI, Ekerljung L, Rådinger M, et al. Cohort profile: the West Sweden Asthma Study (WSAS): a multidisciplinary population-based longitudinal study of asthma, allergy and respiratory conditions in adults. *BMJ open* . 2019;9(6):e027808-e027808. doi:10.1136/bmjopen-2018-027808

10. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* . 2012;28(1):112-118. doi:10.1093/bioinformatics/btr597

11. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology* . 2020;20(1)doi:10.1186/s12874-020-01080-1

12. Wilson S. miceRanger: Multiple Imputation by Chained Equations with Random Forests. https://CRAN.R-project.org/package=miceRanger

13. Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. PMLR; 2016:478-487.

14. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set.*Journal of statistical software* . 2014;61:1-36.

15. John CR, Watson D, Russ D, et al. M3C: Monte Carlo reference-based consensus clustering. *Sci Rep* . 2020;10(1)doi:10.1038/s41598-020-58766-1

16. John CR, Watson D, Russ D, et al. M3C: Monte Carlo reference-based consensus clustering. *Sci Rep* . 2020;10(1):1816.

17. Prosperi MCF, Sahiner UM, Belgrave D, et al. Challenges in Identifying Asthma Subgroups Using Unsupervised Statistical Learning Techniques. Article. *American journal of respiratory and critical care medicine* . Dec 2013;188(11):1303-1312. doi:10.1164/rccm.201304-0694OC

18. Weatherall M, Shirtcliffe P, Travers J, Beasley R. Use of cluster analysis to define COPD phenotypes. Editorial Material. *European Respiratory Journal* . Sep 2010;36(3):472-474. doi:10.1183/09031936.00035210

19. Weatherall M, Travers J, Shirtcliffe PM, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis. Article.*European Respiratory Journal* . Oct 2009;34(4):812-818. doi:10.1183/09031936.00174408

20. Bochenek G, Kuschill-Dziurda J, Szafraniec K, Plutecka H, Szczeklik A, Nizankowska-Mogilnicka E. Certain subphenotypes of aspirin-exacerbated respiratory disease distinguished by latent class analysis. *Journal of Allergy and Clinical Immunology* . 2014 2014;133(1):98-103.

21. Jeong A, Imboden M, Hansen S, et al. Heterogeneity of obesity-asthma association disentangled by latent class analysis, the SAPALDIA cohort.*Respiratory medicine* . 2017 2017;125:25-32.

22. Kaneko Y, Masuko H, Sakamoto T, et al. Asthma phenotypes in japanese adults - their associations with the CCL5 and ADRB2 genotypes.*Allergology International* . 2013 2013;62(1):113-121.

23. Kim MA, Shin SW, Park JS, et al. Clinical characteristics of exacerbation-prone adult asthmatics identified by cluster analysis.*Allergy, Asthma and Immunology Research* . 2017 2017;9(6):483-490.

24. Nadif R, Febrissy M, Andrianjafimasy M, et al. Adult asthma phenotypes identified by a cluster analysis on clinical and biological characteristics. *European Respiratory Journal* . 2018 2018;52:2.

25. Sakagami T, Hasegawa T, Koya T, et al. Identification Of Clinical Asthma Phenotypes By Using Cluster Analysis With Simple Measurable Variables In Japanese Population. *American journal of respiratory and critical care medicine* . 2011 2011;183:1.

26. Loureiro CC, Sa-Couto P, Todo-Bom A, Bousquet J. Cluster analysis in phenotyping a Portuguese population. *Revista Portuguesa de Pneumologia (English Edition)* . 2015 2015;21(6):299-306.

27. Nadif R, Febrissy M, Andrianjafimasy MV, et al. Endotypes identified by cluster analysis in asthmatics and non-asthmatics and their clinical characteristics at follow-up: the case-control EGEA study. *BMJ open respiratory research* . 2020;7(1):e000632.

28. Nagasaki T, Matsumoto H, Kanemitsu Y, et al. Integrating longitudinal information on pulmonary function and inflammation using asthma phenotypes. *Journal of Allergy and Clinical Immunology* . 2014 2014;133(5):1474-U406.

29. Tay TR, Choo XN, Yii A, et al. Asthma phenotypes in a multi-ethnic Asian cohort. *Respiratory medicine* . 2019 2019;157:42-48.

30. Wang L, Liang R, Zhou T, et al. Identification and validation of asthma phenotypes in Chinese population using cluster analysis.*Annals of Allergy Asthma & Immunology* . 2017 2017;119(4):324-332.

31. Seino Y, Hasegawa T, Koya T, et al. A Cluster Analysis of Bronchial Asthma Patients with Depressive Symptoms. *Internal Medicine* . 2018 2018;57(14):1967-1975.

32. Ilmarinen P, Tuomisto LE, Niemela O, Tommola M, Haanpaa J, Kankaanranta H. Cluster Analysis on Longitudinal Data of Patients with Adult-Onset Asthma. *Journal of Allergy and Clinical Immunology-in Practice* . 2017 2017;5(4):967-78.

33. Hsiao HP, Lin MC, Wu CC, Wang CC, Wang TN. Sex-Specific Asthma Phenotypes, Inflammatory Patterns, and Asthma Control in a Cluster Analysis. *Journal of Allergy and Clinical Immunology-in Practice* . 2019 2019;7(2):556-67.

34. Kim JH, Chang HS, Shin SW, Baek DG, Son JH, Park CS, Park JS. Lung function trajectory types in never-smoking adults with asthma: Clinical features and inflammatory patterns. *Allergy, Asthma and Immunology Research* . 2018 2018;10(6):614-627.

35. Watanabe S, Koya T, Hasegawa T, et al. Cluster Analysis Of Uncontrolled Asthma In Japanese Population. *American journal of respiratory and critical care medicine* . 2016 2016;193:1.

36. Dudchenko LS, Savchenko VM. Cluster analysis classification of asthmatic pathologic manifestations during stay at the resort. *Tuberculosis and Lung Diseases* . 2018 2018;96(2):16-21.

37. Boudier A, Curjuric I, Basagana X, et al. Ten-Year Follow-up of Cluster-based Asthma Phenotypes in Adults A Pooled Analysis of Three Cohorts. *American journal of respiratory and critical care medicine* . 2013 2013;188(5):550-560.

38. Liang ZY, Liu LY, Zhao HJ, et al. A Systemic Inflammatory Endotype of Asthma With More Severe Disease Identified by Unbiased Clustering of the Serum Cytokine Profile. *Medicine* . 2016 2016;95(25):7.

39. Loza MJ, Djukanovic R, Chung KF, et al. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. *Respiratory Research* . 2016 2016;17:21.

40. Kim TB, Jang AS, Kwon HS, et al. Identification of asthma clusters in two independent Korean adult asthma cohorts. *European Respiratory Journal* . 2013 2013;41(6):1308-1314.

41. Makikyro EMS, Jaakkola MS, Jaakkola JJK. Subtypes of asthma based on asthma control and severity: a latent class analysis. *Respiratory Research* . 2017 2017;18:11.

42. Zaihra T, Walsh CJ, Ahmed S, et al. Phenotyping of difficult asthma using longitudinal physiological and biomarker measurements reveals significant differences in stability between clusters. *BMC Pulm Med* . 2016 2016;16:8.

43. De Vries R, Dagelet YWF, Spoor P, et al. Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. *European Respiratory Journal* . 2018 2018;51(1):10.

44. Fingleton J, Huang KW, Weatherall M, et al. Phenotypes of symptomatic airways disease in China and New Zealand. *European Respiratory Journal* . 2017 2017;50(6):10.

45. Rootmensen G, van Keimpema A, Zwinderman A, Sterk P. Clinical phenotypes of obstructive airway diseases in an outpatient population. *Journal of Asthma* . 2016 2016;53(10):1026-1032.

46. Zein JG, Erzurum SC. Asthma is Different in Women. *Current Allergy and Asthma Reports* . 2015;15(6)doi:10.1007/s11882-015-0528-y

47. Koike F, Otani Y, Oyama S, et al. Cluster analysis of cough variant asthma using exhaled value of forced oscillation technique. *European Respiratory Journal* . 2018 2018;52:3.

48. Amin K. Relationship between inflammatory cells and structural changes in the lungs of asymptomatic and never smokers: a biopsy study. *Thorax* . 2003;58(2):135-142. doi:10.1136/thorax.58.2.135

49. Honkamäki J, Piirilä P, Hisinger-Mölkänen H, et al. Asthma Remission by Age at Diagnosis and Gender in a Population-Based Study. *The Journal of Allergy and Clinical Immunology: In Practice* . 2021;9(5):1950-1959.e4. doi:10.1016/j.jaip.2020.12.015

Hosted file

`Figure 1.docx` available at https://authorea.com/users/771687/articles/856036-machine-learning-derived-asthma-phenotypes-in-a-representative-swedish-adult-population

Hosted file

Figure 2.docx available at https://authorea.com/users/771687/articles/856036-machine-learning-derived-asthma-phenotypes-in-a-representative-swedish-adult-population

Hosted file

Figure 3.docx available at https://authorea.com/users/771687/articles/856036-machine-learning-derived-asthma-phenotypes-in-a-representative-swedish-adult-population

Hosted file

Table 1.docx available at https://authorea.com/users/771687/articles/856036-machine-learning-derived-asthma-phenotypes-in-a-representative-swedish-adult-population

Hosted file

Table 2.docx available at https://authorea.com/users/771687/articles/856036-machine-learning-derived-asthma-phenotypes-in-a-representative-swedish-adult-population