U.PORTO

FMUP FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO

PHENOTYPES OF CHRONIC DISEASES OF THE AIRWAYS

TOWARDS MULTIDIMENSIONAL DATA-DRIVEN PROFILING

Rita da Silva Amaral

Tese de Doutoramento Apresentada à Faculdade de Medicina da Universidade do Porto em Investigação Clínica e em Serviços de Saúde

Porto, 2019

PHENOTYPES OF CHRONIC DISEASES OF THE AIRWAYS

TOWARDS MULTIDIMENSIONAL DATA-DRIVEN PROFILING

Rita da Silva Amaral

Orientador:

Prof. Doutor João A. Fonseca

MEDCIDS- Departamento de Medicina da Comunidade, Informação e Decisão em Saúde, Faculdade de Medicina da Universidade do Porto

CINTESIS-Centro de Investigação em Tecnologias e Serviços de Saúde, Faculdade de Medicina da Universidade do Porto

Co-Orientador:

Prof. Doutor Tiago Jacinto

CINTESIS-Centro de Investigação em Tecnologias e Serviços de Saúde, Faculdade de Medicina da Universidade do Porto

Escola Superior de Saúde, Instituto Politécnico do Porto

Tese de Doutoramento Apresentada à Faculdade de Medicina da Universidade do Porto em Investigação Clínica e em Serviços de Saúde

Porto, 2019

Art.º 48º, § 3º - **"A Faculdade não responde pelas doutrinas expendidas na dissertação."** (Regulamento da Faculdade de Medicina da Universidade do Porto – Decreto-Lei nº 19337 de 29 de Janeiro de 1931)

Júri da Prova de Doutoramento

Doutor Rui Manuel Lopes Nunes (Presidente)
Professor catedrático da Faculdade de Medicina da Universidade do Porto

Doutor Luís Manuel Taborda Barata
Professor catedrático da Faculdade de Ciências da Saúde da Universidade da Beira Interior

Doutor José Luís Dias Delgado
Professor associado com agregação da Faculdade de Medicina da Universidade do Porto

Doutor Pedro Miguel Carvalho Diogo Carreiro Martins
Professor auxiliar da NOVA Medical School | Faculdade de Ciências Médicas da Universidade Nova de Lisboa

Doutora Maria Cristina Prudêncio Pereira Soares
Professora coordenadora com agregação da Escola Superior de Saúde do Instituto Politécnico do Porto

Doutora Cláudia Catarina Ferreira Chaves Loureiro Barbosa e Silva
Assistente convidada da Faculdade de Medicina da Universidade de Coimbra

Doutor João Almeida Lopes da Fonseca (Orientador)
Professor auxiliar com agregação da Faculdade de Medicina da Universidade do Porto

The work of this PhD thesis was developed at the Center for Health Technology and Services Research (CINTESIS), in Faculty of Medicine of University of Porto, and supported by a PhD grant (reference PD/BD/113659/2015), financed by the European Social Fund and national funds of MCTES (Ministério da Ciência, Tecnologia e Ensino Superior) through FCT (Fundação para a Ciência e Tecnologia, I.P) PhD Programme ref PD/0003/2013 - Doctoral Programme in Clinical and Health Services Research (PDICSS).

To Luís

“The preeminent goal of science is to encompass a maximum of empirical contents through logical deduction with a minimum of hypotheses or axioms.”

Albert Einstein

Acknowledgments

The execution of this thesis would not be possible without an exquisite team work, involving different persons and institutions. To them, I would like to express my sincere gratitude considering their crucial help.

First and foremost, I would like to thank my supervisor, Prof. João A. Fonseca, without whom this thesis would not even start. I am incredibly grateful for his patience, support, encouragement and enthusiasm, which has gone above and beyond expectations. He is, without doubt, an example of academic professionalism and practical scientific knowledge.

I am deeply grateful to my co-supervisor, Prof. Tiago Jacinto, for his scientific assistance, remarkable constructive criticism and intellectual shrewdness and, most of all, for always be a friend. I truly believe I could not have done it without his support and help.

I would also like to thank Prof. Altamiro Costa-Pereira, director of the Faculty of Medicine, University of Porto, former head of the PhD programme in Clinical and Health Services Research, for giving me the chance to be a part of such a fantastic program.

Thank you to Prof. Luís Filipe Azevedo, current head of the PhD program in Clinical and Health Services Research, and to all the professors in this PhD program, for creating an excellent environment to acquire and internalize scientific knowledge and skills.

I am greatly appreciative of all my co-authors for a successful collaboration. I had the privilege to cooperate with very outstanding academic and generous persons that crucially contributed to the development of this work.

Many thanks to the administrative and research staff of the CINTESIS, for all their assistance and patience at answering my numerous queries and providing me an excellent work environment and support. I would like to thank to all “G6 team”, particularly Cristina Jácome and Rute Almeida, for their useful comments and advices throughout these years.

A very special word of thanks goes to Ana Sá e Sousa, an exceptional person that is always present in good and bad moments, for her true friendship through long time. Thank you for getting me through this in one piece.

I would like to thank to all the CUF Porto team, specially Dr. Ana Margarida Pereira, Dr. Luís Araújo and Dr. Ivânia Gonçalves, for the clinical and technical knowledge, and a big thanks for the friendship, inter-aid, and all the times that we spent together.

Many thanks to my father, husband, sisters and all the rest of my family, for their support, but particularly to my mother, the most outstanding and resilient woman I have ever met. She is my biggest inspiration to set high goals for myself and work hard to reach them without disregarding my loved ones.

Finally, to my son Luís, I thank his unconditional love and for teaching me how to be a fighter and never give up even when all the odds are against us. I have learnt from you so much through these 4 years and I am very eager for the following ones.

Table of Contents

Abstract

Background: Profiling of chronic diseases of the airways (CDA) is becoming increasingly relevant to choose the most adequate treatment for each patient. Frequently, CDA have common risk factors and can occur in the same patient. There is a need of moving from a theory-driven (imposed by current knowledge) to an approach that identifies groups of patients (phenotypes) with similar characteristics and response to treatments, in an unsupervised manner (data-driven).

Aim: To explore multidimensional models, supported by advanced statistical methods, for (re)classification of phenotypes of CDA, based on clinical, functional, and immuno-inflammatory characteristics. The specific aims were: 1) describe the proportion of overlap of five commonly reported asthma phenotypes (theory-driven), among adults from the general population and to examine their association with asthma-related outcomes; 2) compare previously defined theory-driven with newly derived data-driven asthma phenotypes, identified by latent class analysis (LCA); 3) identify distinct phenotypes of allergic respiratory diseases obtained by LCA and more comprehensive asthma-related variables, and then, to distinguish each phenotype using classification and regression tree (CART) analysis.

Methods: Data from two independent datasets, derived from general population, were analysed: participants in the United States (US) National Health and Nutrition Examination Surveys (NHANES) from 2007-2012 (n=30,442) and Portuguese participants in the Control and Burden of Asthma and Rhinitis (ICAR), a nationwide cross-sectional study (n=858). First, adults (≥ 18 years) with current asthma from the NHANES were included (n=1,059). Data were weighted for the US population and analyses were stratified by age (< 40 and ≥ 40 years old). Second, LCA was applied to variables commonly used to subdivide asthma, using the same sample of adults with current asthma from the NHANES. LCA models were derived independently according to both age groups. Third, all adults from the ICAR (n=728) that underwent a structured medical interview combined with blood collection, skin prick tests, spirometry with bronchodilation, and exhaled nitric oxide collection were analysed. LCA was applied to 19 variables and the CART algorithm selected the most likely variables distinguishing LCA-classes.

Results: In NHANES data, a substantial overlap of different theory-driven phenotypes was observed both in subjects aged < 40 years (44%) and ≥ 40 years (54%). About

14% of the current asthma patients were “non-classified”. Regardless of phenotype classification, having concomitant phenotypes was significantly associated with FEV_1<LLN (adjusted OR, 95% CI: 3.21, 1.74–5.94) and ≥ 2 controller medications (2.03, 1.16–3.57). LCA identified two data-driven phenotypes among adults with current asthma, for both age groups. The proportions of the theory-driven phenotypes were similar among the two data-driven phenotypes (p>0.05). Class A<40 years (n=285;75%) and Class A≥40 years (n=462;73%), respectively, were characterized by a predominance of highly symptomatic asthma subjects with poor lung function, compared to Class B<40 years (n=94;25%) and Class B≥40 years (n=170;27%). In the other dataset (ICAR study), a six-class model was obtained. Class 1 (25%): nonallergic participants without bronchial or ocular symptoms. Classes 2 (22%) and 3 (11%): nasal and ocular (low levels) symptoms without nasal impairment, mono-sensitized (Class 2) or polysensitized (Class 3). Class 4 (13%): polysensitized participants with high levels of nasal and ocular symptoms, and nasal impairment. Classes 5 (16%) and 6 (14%): high level of nasal, bronchial and ocular symptoms with nasal impairment (non-allergic or polysensitized, respectively). Participants in classes 5 and 6 had more exacerbations and unscheduled medical visits (p<0.001). Ocular symptoms were significantly higher in classes with nasal impairment, compared to those without impairment (p<0.001) or no nasal symptom (p<0.001). CART algorithm highlighted ocular symptoms as the most relevant variable in distinguishing LCA-classes.

Conclusions: 1) A prevalent overlap of commonly reported asthma phenotypes was observed among adults with asthma from the US general population, with implications for objective asthma outcomes. 2) The clinical and physiological variables commonly used to subdivide asthma seem to be insufficient to differentiate specific asthma phenotypes among these population, irrespective of using data-driven or theory-driven approaches. 3) Applying a more comprehensive disease features available in the ICAR study, revealed novel severe phenotypes with co-occurrence of ocular, nasal and bronchial symptoms, and prone to exacerbations.

In summary, the complexity and unique features of phenotyping CDA requires a combination of unsupervised analysis (data-driven) and clinical knowledge with broader data availability, to provide a better taxonomy of these conditions.

Resumo

Introdução: A classificação das doenças crónicas das vias aéreas tem-se tornado cada vez mais importante na escolha do tratamento mais adequado para cada indivíduo. Frequentemente, estas doenças têm fatores de risco comuns e podem ocorrer simultaneamente no mesmo doente. Existe uma necessidade de mudar de uma abordagem orientada pela teoria (imposta pelo conhecimento atual, *theory-driven*) para uma abordagem que identifique grupos de doentes (fenótipos) com características e respostas terapêuticas semelhantes, de modo não-supervisionado (orientada por dados, *data-driven*).

Objetivos: Explorar modelos multidimensionais, apoiados em métodos estatísticos avançados, para (re)classificar fenótipos de doenças crônicas das vias aéreas, com base nas características clínicas, funcionais e imuno-inflamatórias. Os objetivos específicos foram: 1) descrever a proporção de sobreposição entre cinco fenótipos de asma habitualmente reportados na literatura (*theory-driven*), em adultos da população geral e avaliar a sua influência nas manifestações da asma; 2) comparar os fenótipos de asma previamente definidos com novos fenótipos *data-driven*, identificados pela análise de classes latentes (LCA); 3) identificar fenótipos distintos de doenças respiratórias alérgicas obtidos por LCA, utilizando um conjunto abrangente de variáveis e distinguir cada fenótipo através da análise de classificação e árvore de regressão (CART).

Métodos: Foram analisados dados de duas bases de dados independentes, obtidos da população geral: participantes no inquérito *National Health and Nutrition Examination Surveys* (NHANES) dos Estados Unidos da América (EUA), entre 2007-2012 (n=30,442) e participantes no estudo nacional e transversal Impacto e Controlo da Asma e Rinite (ICAR), realizado em Portugal (n=858). Primeiro, foram incluídos adultos (≥18 anos) com asma do NHANES (n=1,059). Os dados foram ponderados para a população dos EUA e a análise foi estratificada por idade (< 40 e ≥ 40 anos). A LCA foi, então, aplicada às variáveis habitualmente utilizadas para subdividir a asma, usando a mesma amostra de adultos com asma do NHANES. Os modelos de LCA foram obtidos de forma independente de acordo com os dois grupos etários. Por fim, foram analisados dados de todos adultos do ICAR (n=728), submetidos a uma entrevista médica estruturada com colheita de sangue, testes cutâneos, espirometria com prova de broncodilatação e medição de óxido nítrico exalado. A LCA foi aplicada

a 19 variáveis e o algoritmo CART selecionou as mais úteis para a diferenciação das classes obtidas por LCA.

Principais resultados: Nos dados do NHANES, foi observada uma elevada sobreposição entre os fenótipos de asma definidos pela abordagem *theory-driven*, na população geral, em indivíduos com <40 anos (44%) e ≥ 40 anos (54%). Cerca de 14% dos pacientes com asma não eram enquadráveis em nenhum dos fenótipos conhecidos. Independentemente da classificação fenotípica, ser incluído em mais de um fenótipo foi significativamente associado a FEV_1<LLN (OR ajustado, IC 95%: 3.21, 1.74–5.94) e ter ≥2 medicamentos de controlo (2.03, 1.16–3.57). A LCA identificou dois fenótipos *data-driven* em adultos com asma, para ambos os grupos etários. As proporções dos fenótipos *theory-driven* foram semelhantes entre os dois fenótipos *data-driven* (p> 0.05). Classe A<40 anos (n=285;75%) e Classe A≥40 anos (n=462;73%), foram caracterizados pela predominância de indivíduos muito sintomáticos e com má função pulmonar, comparativamente à classe B < 40 anos (n=94;25%) e classe B≥40 anos (n=170;27%), respetivamente. No outro conjunto de dados (estudo ICAR), foi obtido um modelo de seis classes. Classe 1 (25%): participantes não-alérgicos sem sintomas brônquicos nem oculares. Classes 2 (22%) e 3 (11%): sintomas nasais e oculares (níveis baixos) sem limitações nasais, mono-sensibilizados (Classe 2) ou polissensibilizados (Classe 3). Classe 4 (13%): indivíduos polissensibilizados com elevada sintomatologia nasal e ocular e com limitações nasais. Classes 5 (16%) e 6 (14%): elevada sintomatologia nasal, brônquica e ocular, com comprometimento nasal (não-alérgicos ou polissensibilizados, respetivamente). Os indivíduos das classes 5 e 6 tiveram mais exacerbações e consultas médicas não programadas (p<0,001). Os sintomas oculares foram significativamente mais elevados nas classes com comprometimento nasal, em comparação com indivíduos sem comprometimento (p<0,001) ou sintomatologia nasal (p<0,001). O algoritmo CART evidenciou os sintomas oculares como a variável mais relevante na diferenciação das classes de LCA.

Conclusões: 1) Foi observada uma elevada sobreposição entre os fenótipos de asma habitualmente reportados na literatura, em adultos com asma da população geral dos EUA, com implicações objetivas na asma. 2) As variáveis clínicas e fisiológicas normalmente utilizadas para subdividir a asma parecem ser insuficientes na diferenciação de fenótipos de asma específicos nesta população, independentemente da utilização de abordagens *theory-driven* ou *data-driven*. 3) A aplicação de características da doença mais abrangentes e disponíveis no estudo ICAR, revelou

novos fenótipos graves, com sintomas oculares, nasais e brônquicos concomitantes, e suscetíveis a exacerbações.

A complexidade e características ímpares da classificação das doenças crónicas das vias aéreas requerem uma combinação entre análise não-supervisionada (*data-driven*) e conhecimento clínico, com uma maior acessibilidade de dados, para auxiliar uma melhor taxonomia destas patologias.

Abbreviations

ACOS	Asthma-COPD Overlap Syndrome
AIC	Akaike Information Criterion
AR	Allergic Rhinitis
ARD	Allergic Respiratory Disease
ARIA	Allergic Rhinitis and its Impact on Asthma
ATS	American Thoracic Society
AwObesity	Asthma with Obesity
AwCOPD	Asthma with concurrent COPD
B-Eos	Blood Eosinophils count
BIC	Bayesian Information Criterion
BMI	Body Mass Index
BOLD	Burden of Obstructive Lung Disease
CART	Classification and Regression Tree
CDA	Chronic Diseases of the Airways
CI	Confidence Interval
COPD	Chronic Obstructive Pulmonary Disease
CRD	Chronic Respiratory Diseases
ECP	Eosinophilic Cationic Protein
ED	Emergency Department
ERS	European Respiratory Society
FeNO	Fraction of exhaled Nitric Oxide
FEV_1	Forced Expiratory Volume in one second
FVC	Forced Vital Capacity
GARD	Global Alliance against chronic Respiratory Diseases
ICAR	Control and Burden of Asthma and Rhinitis
ICS	Inhaled Corticosteroids
IgE	Immunoglobulin E
IL	Interleukin
INAsma	Portuguese National Asthma Survey
LABA	Long-acting inhaled β2-agonist
LCA	Latent Class Analysis
LLN	Lower Limit of Normal
MCA	Multiple Correspondence Analysis
MeDALL	Mechanisms of the Development of ALLergy

mHealth	Mobile Health
NAR	Nonallergic Rhinitis
NHANES	National Health and Nutrition Examination Survey
OCS	Oral corticosteroids
PAM	Partition around medoids
ppb	parts per billion
RAKE	Rapid Automatic Keyword Extraction
SPT	Skin prick test
Th2	T-helper cell type 2
US	United States
WHO	World Health Organization

List of Publications

The following journal articles have been published as original research based on the work included in this thesis (Art.º 8º do Decreto-Lei n.º 388/70). The full-text of the papers can be found in the Appendices section of this thesis.

Study I. Amaral, R., Fonseca, J. A., Jacinto, T., Pereira, A. M., Malinovschi, A., Janson, C., & Alving, K. (2018). Having concomitant asthma phenotypes is common and independently relates to poor lung function in NHANES 2007–2012. *Clinical and Translational Allergy*, *8*:13. doi:10.1186/s13601-018-0201-3.

[2017 Journal Impact Factor: 3.54 (Journal Citation Reports®); 2nd Quartile of "Immunology and Allergy" (No. 47/190)]

Study II. Amaral, R., Pereira, A. M., Jacinto, T., Malinovschi, A., Janson, C., Alving, K., & Fonseca, J.A. (2019). Comparison of hypothesis- and data-driven asthma phenotypes in NHANES 2007-2012: the importance of comprehensive data availability. *Clinical and Translational Allergy, 9*:17. doi.org/10.1186/s13601-019-0258-7.

[2017 Journal Impact Factor: 3.54 (Journal Citation Reports®); 2nd Quartile of "Immunology and Allergy" (No. 47/190)]

Study III. Amaral, R., Bousquet, J., Pereira, A.M., Araújo, L., Sá-Sousa, A., Jacinto, T., Almeida, R., Delgado, L. & Fonseca, J.A. (2019). Disentangling the heterogeneity of allergic respiratory diseases by latent class analysis reveals novel phenotypes". *Allergy, 74*(4), 698-708. doi: 10.1111/all.1367 0. Epub 2018 Dec 5.

[2017 Journal Impact Factor: 6.05 (Journal Citation Reports®); 1st Quartile of "Immunology and Allergy" (No. 21/190) and "Immunology" (n. 29/209)]

Additionally, during the PhD work, the following related abstract publications were presented at international conferences:

- Amaral, R., Pereira, A.M., Jacinto, T., Malinovschi, A., Janson, C., Alving, K., & Fonseca, J.A. "Identification of asthma phenotypes in the US general population: a latent class analysis approach". - ***Oral communication*** *and poster presentation at European Academy of Allergy and Clinical Immunology (EAACI) International Congress 2019, Lisbon, Portugal.*

- Amaral, R., Jacinto, T., Sousa-Pinto, B., & Fonseca, J.A. “Where do we stand with asthma phenotypes derived from data-driven methods? A systematic review”. - *Poster presentation at European Respiratory Society (ERS) International Congress 2019, Madrid, Spain.*

- Amaral, R., Jacinto, T., Pereira, A., Almeida, R., & Fonseca, J. (2018). A comparison of unsupervised methods based on dichotomous data to identify clusters of airways symptoms: latent class analysis and partitioning around medoids methods. *European Respiratory Journal, 52*(Suppl. 62, PA4429). - *Poster presentation at ERS International Congress, September 2018, Paris, France.*

- Amaral, R., Pereira, A.M., Araújo, L., Sá-Sousa, A., Jacinto, T., Almeida, R. … Fonseca, J.A. (2018). Phenotyping allergic respiratory diseases: An unsupervised classification using latent class analysis. *Allergy,73*(S105) - *Poster presentation at EAACI International Congress, May 2018, Munich, Germany* – ***Prize for outstanding poster presentation.***

- Amaral, R., Fonseca, J.A., Jacinto, T., Malinovschi, A., Janson, C., & Alving, K. (2016). High proportion of overlap between adult asthma phenotypes in a large population-based sample. *European Respiratory Journal,48*(Suppl. 60, PA568). *Poster presentation at ERS International Congress, September 2016, London, UK.*

1. Introduction

This chapter is an overview of the paradigm of precision medicine and of phenotyping the chronic diseases of the airways, that are relevant for the remainder of this thesis. It is divided in four main topics: 1) Precision medicine, focusing on the shift of medical paradigms; 2) Data mining applied to healthcare data; 3) Phenotypes, their definitions and data-driven techniques; and 4) Dimensions of the phenotypes of chronic diseases of the airways, focusing data-driven asthma phenotypes.

1.1 Precision Medicine

As early as the year 370 BC, Hippocrates, the father of Western medicine, famously said that "it is far more important to know what person the disease has than what disease the person has". Today, it is clear the urgent need to shift the old paradigm in medicine of diagnosing and classifying the diseases into organ systems and specialized field, towards a tailored prevention, diagnostic and treatment of the individual.

However, most of the current healthcare systems still operate with the “one-size-fits-all” approach, treating the disease or symptoms instead of the person (Vanfleteren et al., 2014). For this reason, treatments can be very successful for some patients but not for others, leading to poor outcomes, unnecessary suffering and elevated direct and indirect costs. In 2011, the Institute of Medicine used the term “Precision Medicine” as an emerging approach for disease and people stratification based on an individual’s genetic, environment and lifestyle variability (National Research Council, 2011).

This new comprehensive paradigm in medicine aims at the patient’s stratification of clinical and functional heterogeneity of symptom profiles, conditions and responses to therapy, by identifying patient-to-patient variation. Based on the Precision Medicine approach, a preventive or therapeutic intervention targeting individual characteristics can allow a more effective and personalized approach to patient care (Bousquet et al., 2016; Hodson, 2016).

Before the term Precision Medicine, a related concept was called “P4 medicine” (predictive, preventive, personalised and participative), which can be summarized as the convergence of three main components: 1) systems biology approaches, 2) analytics of big data and 3) the patient-driven health management (Galas & Hood,

2009). The personalization aspect is addressed by the Precision Medicine, that formalizes “a framework for developing a more precise and more accurate classification of disease based on molecular biology” (National Research Council, 2011). This “New Taxonomy” of disease - also known as redefining the disease phenotype - could not only lead to tailored treatments, but also to individualized prevention strategies and personalized diagnosis (Flores, Glusman, Brogaard, Price, & Hood, 2013).

Precision Medicine has been increasingly recognized as the way forward for optimizing patient care (Hellings et al., 2017). However, major collaborations between interdisciplinary fields are required to integrate different sources of information, and to gain a comprehensive understanding of biology and medicine. Due to the large sizes and complex nature of biomedical systems, data integration remains a challenge in applying the Precision Medicine concept to omics data types and clinical datasets of patient features (Gligorijević, Malod-Dognin, & Pržulj, 2016). Modelling complex biological systems requires linking knowledge across many levels of science, from genes to disease, to patient and environment. Further, the data characteristics of the problems have also grown from static to dynamic and spatiotemporal, complete to incomplete, and centralized to distributed, and grow in their scope and size - this is known as “Big Data” (Chu, 2014).

The term “Big Data” refers not only to the availability of large volumes of healthcare data, but also to the complexity of this data, typically seen in genetic, environmental and phenotypic data. It is often characterized by three Vs (volume, velocity and variety), where traditional databases and/or processing methods are inadequate or insufficient (De Mauro, Greco, & Grimaldi, 2016). Moreover, Big Data provides the potential for “learning” patterns or predicting health outcomes and optimal treatment strategies based on prior information (Belgrave et al., 2016).

Using Big data could potentially help in a more personalized and precision medicine for patients by improving diagnosis’ accuracy, and therapy tailored to the individual (Gligorijević et al., 2016). This promise comes from data collected from numerous sources, ranging from molecules to individuals and populations, and the integration of these data into networks that improve understanding of heath and disease (Asri, Mousannif, Al Moatassime, & Noel, 2015).

The evolution of powerful tools and technologies to analyse such high-dimensional, large datasets has expanded the traditional and accepted disease phenotypes. The

process of exploring data in search of consistent patterns and/or systematic relationships between variables is called “Data Mining” (Miralles et al., 2014).

1.2 Data Mining in Healthcare

The past decade has seen an exponential growth in the use of genomics, proteomics and functional genomics in biomedical research. Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Han, Kamber & Pei, 2012). Data Mining has been successfully applied to diverse areas, given its potential for solving complex problems (Basile & Ritchie, 2018). Many of these applications search for patterns in complex structural information.

Despite the concept of Data Mining going back to 1937, when Alan Turing introduced the idea of an “Universal Machine”, that could perform computations similar to those of modern-day computers (Turing, 1937), this is considered a relatively new interdisciplinary field. Data Mining comprises areas such as database systems, data warehousing, statistics/Machine Learning, data visualization, information retrieval, and high-performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, and signal processing (Han et al., 2012).

Machine Learning provides the technical basis of Data Mining. It can look at patterns and learn from them to adapt future behaviours, using Data Mining as an information source. Particularly, the essential role of statistics within Machine Learning in the context of Big Data analysis is now often called Statistical Learning (Bednekoff, 2008). Statistical Learning emphasizes models and their interpretability, precision and uncertainty, while Machine Learning has a greater emphasis on large scale applications and prediction accuracy, such as Marketing (Hastie, Tibshirani, & Friedman, 2009).

The methods that underlie Statistical Learning are those who understand “what the data says” - learning from data, i.e. methods able to explore and retain structures/patterns from data that is replicable across different samples extracted from the same population (Hastie et al., 2009). There are three main categories of learning from data (Everitt, Landau, Leese, & Stahl, 2011; Huddleston & Brown, 2018):

- *Supervised learning*, which typically involves building an algorithm in which the input is a dataset of predictors - known as features or attributes (e.g. age, biomarkers,

lung function) and can predict the value of a specific outcome or output (e.g. asthma exacerbations) (Huddleston & Brown, 2018). Supervised learning includes classification (categorical outcome) and regression problems (quantitative outcomes). The first includes the following approaches: 1) Mathematical formulae (e.g. linear discriminants); 2) Logical approaches (e.g. classification trees); 3) Probabilistic (e.g. naive Bayes); and 4) others (e.g. neural networks and supporting vector machines). Regression problems include: 1) Mathematical formulae (e.g. linear regression, multiple adaptive regression splines, etc.); 2) Logical approaches (e.g. regression trees); and 3) others (e.g. neural networks and supporting vector machines) (T. Hastie et al., 2009). These different approaches entail different compromises in terms of prediction error, computational complexity and model interpretability (Prosperi et al., 2013).

- *Unsupervised learning*, where there is no predefined outcome to be predicted, just a set of predictors measured on a set of samples. The goal is to find groups of samples/features that behave similarly (e.g. groups of patients who share similar clinical or test result profiles) (Huddleston & Brown, 2018). Moreover, unsupervised learning can find linear combinations of features with the most variation, discovering the data structure. The two foundations of this method are 1) clustering, that can be either distance-based i.e. distance/similarity between observations (e.g. Agglomerative/Divisive Hierarchical Clustering, Partition around medoids (PAM) and Fuzzy Clustering) or model-based i.e. a based on a probability model for the data (e.g. latent class analysis); and 2) dimensionality reduction (e.g. principal components analysis and factor analysis) (Hastie et al., 2009).

- *Semi-Supervised learning*, that combines insights from both supervised and unsupervised methods by exploring observations where the outcome is known only for a small amount of data (e.g. build an algorithm using data of patients' profile that response positively or negatively to a drug and then training the algorithm in patients with unknown treatment outcome) (Huddleston & Brown, 2018).

The vast increase in huge volumes of high-dimensional, and heterogenous healthcare data lead to a shift in "formulate hypothesis, build model, and evaluate results" paradigm (traditional hypothesis-based research) toward a data-driven hypothesis generating paradigm, with the process: "collect and store data, mine for new hypotheses, confirm with data or experimentation" (Han et al., 2012). This latter approach is an advantage in the case of heterogeneous diseases with possibly

different underlying pathophysiological mechanisms, such as asthma or other chronic diseases of the airways. Moreover, it is aligned with the concept of Precision Medicine, as it seeks to identify patterns of the disease through deep phenotyping (disease subclassification), allowing for optimal treatment based on an individual's unique combination of genes, environment, and comorbidities (Ashley, 2016).

The need to refocus efforts to propose a new taxonomy of airway diseases is emphasised by the emergence of highly specific therapies, since a positive response is more likely to be phenotype-specific rather than disease-specific (Bafadhel et al., 2011; Pavord et al., 2018). Therefore, a new era of airways disease phenotyping has emerged, incorporating the Precision Medicine principles and Big Data analytics (Statistical Learning) into daily care of patients.

1.3 Phenotyping

One of the first attempts to establish a system for disease classification, largely based on symptoms, was undertaken by the renowned taxonomist Carl Linnaeus, in his *Genera Morborum* (Varieties of Diseases) (Linnaeus, 1759). The rise of data-intensive biology, advances in information technology, and changes in the way healthcare is delivered have created the urgent need to create a "New Taxonomy" (National Research Council, 2011), novel phenotypes, that integrates multidimensional data (Wardlaw et al., 2005).

The classic definition of phenotype is "the observable characteristics without direct relation with the pathological process" (Rice, Saccone, & Rasmussen, 2001). A clinical phenotype is "a single or combination of disease attributes that describe differences between individuals as they relate to clinically meaningful outcomes, such as symptoms, exacerbations, response to therapy, rate of disease progression, or death" (Robinson, 2012). Furthermore, disease phenotyping stratifies a heterogeneous group of patients with a specific disease into homogeneous subgroups, based on clinical, molecular, or other types of patient features (Delude, 2015).

Accurate phenotyping and disease stratification into subtypes according to their underlying biological mechanisms are fundamental steps towards Precision Medicine. However, they often do not capture the full diversity of clinical and even pathophysiological manifestations (Delude, 2015). Disease stratification using classical diagnostic methods that relies solely on signs and symptoms may not be sufficient for a more effective and personalized patient care (Miralles et al., 2014), particularly in those

patients with co-occurrence of two or more chronic medical conditions, such as the asthma-rhinitis phenotype (Siroux et al., 2018).

The interaction of different diseases and the impact they have on clinical outcomes must be considered in daily clinical practice – the multimorbidity concept (Barnett et al., 2012). Multimorbidity, therefore, constitutes a broader, patient-centred concept, in contrast to comorbidity that is an index disease-based concept (Valderas, Starfield, Sibbald, Salisbury, & Roland, 2009). Multimorbidity considers all aspects of a patient's condition, including the potential disease interaction and potential pathophysiological links; however, it is complex and may be difficult to manage in clinical practice (Barnett et al., 2012).

1.3.1 Theory- and Data-driven classification

Current approaches used to classify patients into sub-groups of diseases are: theory-driven (imposed by current knowledge) and data-driven (hypothesis-generating). In the first approach, diseases are subclassified *a priori* either by the presence or absence of specific risk factors or based on clinical and/or molecular phenotypes. Conversely, in the data-driven approach, no prior disease classification is required (Bousquet et al., 2011b; Prosperi et al., 2013) and often start with a broad hypothesis and using (or collecting) data relevant to that hypothesis. Then, these data are explored to generate more specific and automatic hypotheses, providing new insights into phenotypes of complex disease pathogenesis and novel reclassifications (Bousquet et al., 2011b).

Airways diseases do not escape the great steps of unravelling their complexity through the refinement of the clinical phenotypes using of data-driven techniques. Recently, a combination of both hypothesis- and data-driven approaches was proposed to assess multimorbidity of allergic diseases of the airways by Mechanisms of the Development of ALLergy, FP7 (MeDALL), using Machine Learning tools (Anto et al., 2017).

1.3.2 Data-driven methods for phenotype classification

Various classes of data-driven (unsupervised) algorithms have been implicated in tackling the problems of traits heterogeneity. This topic will specifically focus on distance-based (clustering analysis) and model-based (latent class analysis) approaches, as they are the most utilized techniques to address phenotypic heterogeneity in healthcare data (Basile & Ritchie, 2018). Moreover, these two

approaches are not mutually exclusive and need to be thoroughly evaluated according to the study and patients' characteristics.

Distance-based approaches

Cluster analysis is a generic name for a wide variety of procedures to identify clusters, homogenous groups of objects (or cases, observations), being very dissimilar to objects not belonging to that cluster (Everitt et al., 2011).

Most clustering methods use the information on the distances among observations in a data set to decide on the natural groupings of the cases using a measure of (dis)similarity to generate a (dis)similarity matrix, with most commonly used being the Euclidean Distance Function (Caillez & Kuntz, 1996) (Figure 1), defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

Figure 1. Formula for the calculation of the Euclidian distance. Where x and y are two observations described by p variables.

The main types of clustering methods are 1) hierarchical, 2) partitional and 3) two-step clustering, which is largely a combination of the first two methods. The hierarchical method generates a hierarchy of groups, from 1 to *n* groups, where *n* is the number of lines in the data set. Hierarchical clustering techniques may be further divided into i) agglomerative methods, which proceed by a series of successive fusions of the *n* individuals into groups, and ii) divisive methods, which separate the *n* individuals successively into finer groupings (Everitt et al., 2011). The main characteristic of these methods is to present the results in a dendrogram, a mathematical and pictorial representation of the complete clustering procedure. The nodes of the dendrogram represent clusters, and the lengths of the stems (heights) represent the distances at which clusters are joined (Everitt et al., 2011). An example is shown in Figure 2.

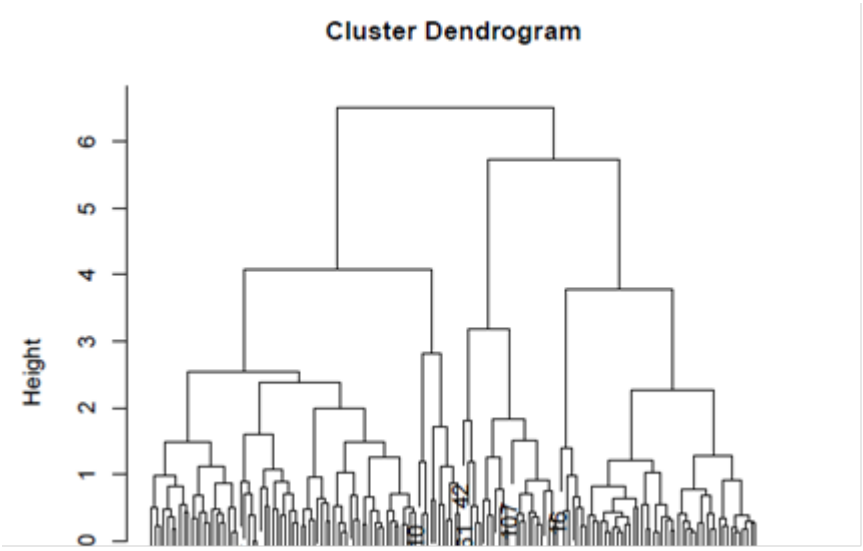


Figure 2. Example of a dendrogram of the hierarchical binary cluster tree. The different possibilities of clustering the data are represented.

The links between objects are represented as upside-down U-shaped lines (Figure 2). The height of the “U” indicates the distance between the objects. In the top, the observations are all in one group and in the bottom, there are as many groups as observations.

The partitional methods divide the data into non-overlapping subsets so that each subject is classified into exactly one subgroup. The most popular methods are k-means and k-medoids (e.g. PAM). K-means clustering seeks the minimum within-cluster variation as a measure to form homogenous clusters and is suitable for quantitative variables (Jain, 2010). Meanwhile, PAM algorithm searches for k representative medoids – the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal – and constructs k clusters by assigning each object to its nearest medoid (Everitt et al., 2011).

Compared to k-means, PAM is more robust to the presence of outliers because it uses original objects as centroids instead of averages that may be subject to the effects of outliers and is used for categorical/binary variables. Moreover, PAM uses a more robust measure of the clustering quality, an absolute error instead of the squared error used in k-means (Kaufman & Rousseeuw, 1990).

Model-based approaches

In model-based clustering methods, the “true” clusters are defined by parametric probability distributions that can be interpreted to generate homogeneous points, and the whole data set is modelled by a mixture of such distributions (Banfield & Raftery, 1993).

This approach assumes a formal statistical model that assumes that the population consists of several subpopulations (clusters) in each the variables have a different multivariate probability density function, resulting in what is known as a finite mixture density. Model-based methods estimate clusters based on the maximum value of the following estimated posterior probability (Everitt et al., 2011). Moreover, this approach is appropriate to variables that have a multivariate normal distribution.

To provide suitable models for categorical data, the mixture modelling uses multivariate Bernoulli densities which arise from assuming that, within each group, the categorical variables are independent of one another, the so-called conditional independence assumption. It is this approach which is the basis of latent class analysis (Goodman, 2009).

Latent class analysis (LCA) classify individuals into homogenous groups (latent classes) using observed response patterns of individuals across a set of categorical (nominal or ordinal) variables (such as symptoms present vs. absent; questionnaire items measured on a Likert scale), based on their maximum likelihood class membership (Collins & Lanza, 2009).

The latent classes divide individuals into mutually exclusive groups, i.e. the individual differences in observed item response patterns are explained by differences in latent class membership, where each class shows a characteristic, class-specific response profile (Oberski, 2016), as graphically described in Figure 3.

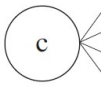


Figure 3. Path diagram for basic latent class analysis model. Where c is a categorical latent variable and u_x are the latent class indicators.

In the diagram, the arrows from c to the latent class indicators u1, u2, u3, and u4 indicate that the thresholds of the latent class indicators vary across the classes of c (Figure 3). This implies that the probabilities of the latent class indicators vary across the classes of c. The arrows correspond to the regressions of the latent class indicators on a set of dummy variables representing the categories of c.

Both distance- and model-based approaches have identified relevant and important features about disease traits, not yet known by expert knowledge (Burgel, Paillasseur, & Roche, 2014; Green et al., 2016). However, each model has strengths and weaknesses, and both will need to be thoroughly evaluated for the application of identifying and refining phenotypes in biomedical data.

Latent class-based methods have been extensively used in a wide scopes of healthcare research (Larsen, Pedersen, Friis, Glümer, & Lasgaard, 2017; Leventhal, Huh, & Dunton, 2014; Rovner, Vowles, Gerdle, & Gillanders, 2015), and also in the identification of phenotypes of chronic diseases of the airways (Bochenek et al., 2014; Cecere et al., 2012; Couto et al., 2015; Jeong et al., 2017). Moreover, simulation results showed that the model-based approach produces substantially less biased estimates of the effect compared to other classification technique (Lanza, Tan, & Bray, 2013).

1.4 Phenotypes of Chronic Diseases of the Airways

1.4.1 Chronic Diseases of the Airways

Airway diseases represent one-sixth of all deaths worldwide and one in eight of all deaths in the European Union (Loddenkemper, Gibson, Sibille, & Lundbäck, 2013). Five respiratory conditions, known as “The Big Five”, include: 1) chronic obstructive pulmonary disease (COPD); 2) asthma; 3) acute respiratory infections; 4) tuberculosis; and 5) lung cancer, which make the largest contribution to morbidity and mortality, and account for most of the burden associated with respiratory diseases (Marciniuk et al., 2014). More than 1 billion people in the world are burdened with chronic respiratory diseases (CRD), which is one among the four major noncommunicable diseases, preventable and treatable diseases (WHO, 2007).

CRDs are a group of chronic diseases affecting the airways and the other structures of the lungs, in which interactions between genetic and environmental lead to harmful inflammatory responses (Beaglehole, Ebrahim, Reddy, Voûte, & Leeder, 2007). The most common and prevalent CRDs worldwide are: asthma (≈ 300 million), COPD (≈ 210 million), allergic rhinitis (AR) (≈ 400 million), and sleep apnoea syndrome (>100 million) (WHO, 2008). In Portugal, respiratory diseases are a major cause of morbidity and mortality, in particular CRDs, whose prevalence is about 40%, with a tendency to increase (Fonseca-Antunes, Bárbara, & Melo-Gomes, 2013).

The presence of CRD is an important risk factor for lower quality of life, particularly among older adults (Carreiro-Martins et al., 2016), possibly due to their comorbidities and higher predisposition to respiratory infections compared to younger subjects (Bentayeb et al., 2013). To raise the recognition of the importance of CRDs as one of the most important health problems globally, an alliance of national and international organizations and institutions supported by the World Health Organization (WHO), was created, the Global Alliance against chronic Respiratory Diseases (GARD) (Bousquet, Dahl, & Khaltaev, 2007; WHO, 2007). In this context, since 2007, GARD is implemented in Portugal, developing activities for the improvement of CRD outcomes and against its burden, at country level (Rosado-Pinto & Carreiro-Martins, 2017). Currently, GARD Portugal is integrated in the Portuguese National Programme for Respiratory Diseases, under the responsibility of the Portuguese Ministry of Health (Fonseca-Antunes et al., 2013).

Among each specific CRD, the ones that have the higher burden and social impact are asthma, AR and COPD (Lozano et al., 2012; Nunes, Pereira, & Morais-Almeida, 2017; Wise et al., 2018), also they often overlap in the same person (Bousquet et al., 2008; Cruz et al., 2007; Kim et al., 2018). The understanding of the overlapping features in the chronic diseases of the airways (CDA) remains incomplete (Bateman et al., 2015).

Overlapping conditions

Each CRD is heterogenous and characterized by defined symptoms, exposure to risk factors, inflammation or patterns of airflow obstruction and airway hyperresponsiveness (Wardlaw, Silverman, Siva, Pavord, & Green, 2005). However, it remains unclear whether all these different phenotypes represent the expression of one single disease with multiple mechanisms or whether some phenotypes represent distinct diseases with similar symptomatology (Pavord et al., 2018).

The CRDs' clinical profile is extensive. Patients may have similar symptoms, common environmental risk factors and frequently they occur in the same patient (Bousquet et al., 2008), originating concepts that aggregate asthma with COPD or asthma with rhinitis (Bousquet et al., 2008; Gibson & Simpson, 2009). In fact, they can be viewed as partially overlapping syndromes, that require a better characterization of the phenotypes found in the general population (Bateman, Reddel, van Zyl-Smit, & Agusti, 2015; Wenzel, 2006).

A vast majority of patients with asthma have concomitant rhinitis (Bousquet et al., 2008). On the other hand, around 20-40% of patients with AR have bronchial symptoms (Bousquet et al., 2008), regardless of having allergic sensitization (Leynaert et al., 2004). The significant overlap between asthma and rhinitis led to the Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines recommending that patients with AR should be tested for asthma (Demoly & Bousquet, 2008).

Regarding asthma and COPD, the most common symptoms are similar, such as dyspnoea, cough, wheezing. The changes in lung function can also be remarkably similar, since reversible airflow limitation, the defining key feature of asthma, may also be observed in some subjects with COPD (Albert et al., 2012). Likewise, an accelerated decline in lung function, a supposedly defining key feature of COPD, has been seen in smoking subjects with asthma (James et al., 2005) and in patients with uncontrolled asthma and with exacerbations (O'Byrne, Pedersen, Lamm, Tan, & Busse, 2009).

Surprisingly, both these respiratory conditions are still diagnosed and treated based on airways symptoms and traditional lung function measures, which is been demonstrated as being an outdated and non-specific approach (Reddel et al., 2015; Vogelmeier et al., 2017), potentially leading to sub-optimal treatments. Furthermore, although it is currently widely accepted that these chronic conditions are complex and include a heterogeneous broad of disease subtypes with different underlying pathophysiological mechanisms (Pavord et al., 2018; Vanfleteren et al., 2014), there is an ongoing debate between those who considered asthma and COPD different expressions of a single disease (the lumpers), concept known as the “Dutch Hypothesis”, and those who favoured splitting into separate entities (the splitters), the “British Hypothesis” (Orie et al., 1961; Ghebre et al., 2015).

Recently, the concept of Precision Medicine has been proposed in the context of CDA based on the integrated assessment of the complex clinical and biological status of individual patients – “Treatable traits” (Agusti et al., 2016). This label-free approach proposes a deconstruction of each CRD into single components, starting with the diagnosis of airways disease, rather than attempting to diagnose a specific CRD, which may not be possible or necessary in clinical practice (Agusti et al., 2016). Treatable traits are “therapeutic targets identified by phenotype or endotype discovery through validated biomarker(s)” (König, Fuchs, Hansen, von Mutius, & Kopp, 2017), and they can coexist and change over time in the same patients, surpassing the phenotype concept (Agustí et al., 2017). This is considered a promising strategy to choose the most appropriate therapeutic strategy for individual patients, and to provide optimal improvement of disease control and quality of life (Agustí et al., 2015).

Therefore, there is an unprecedented potential to go beyond the simplistic concept that ignores individual heterogeneity. Currently, there are major research opportunities due to the recent development and establishment of techniques for sample collection from the airways and new biostatistical methods for integrating data from multiple sources and levels (Agustí et al., 2017). Thus, it is an ideal time to re-examine the identification of novel CRDs phenotypes using detailed information from patients.

Rhinitis

Rhinitis is characterized by inflammation of the nasal mucosa causing nasal obstruction, runny nose, sneezing, and/or pruritus in the nose (Greiner, Hellings, Rotiroti, & Scadding, 2011), affecting 23% of European adults (Akdis et al., 2015). In

Portugal, the overall estimated prevalence ranges from 26.1% to 43.4% (Morais-Almeida et al., 2013a; Morais-Almeida et al., 2013b; Todo-Bom et al., 2007).

Three main phenotypes of rhinitis are distinguishable based on history and clinical examination: AR, infectious rhinitis, and nonallergic non-infectious rhinitis (NAR) (Papadopoulos et al., 2015). If we disregard infectious rhinitis, the most common rhinitis phenotype is AR, in which a clinical response to an otherwise innocent environmental factor or allergen in combination with specific immunoglobulin E (IgE) targeting aeroallergens results in symptoms (Greiner et al., 2011). AR is usually subdivided in different phenotypes based on duration (intermittent/persistent) and severity of disease (mild/moderate-severe, and severe chronic upper airways disease), clinical presentation with most bothersome symptoms (“runners”, “sneezers”, “blockers”), presence of comorbidities (such as asthma, sinusitis, conjunctivitis), response to treatment (with or without corticoid response), time trend (seasonal, perennial) (Droessaert et al., 2016; Papadopoulos et al., 2015), and level of control (evaluated by several tools such as Control of Allergic Rhinitis and Asthma Test (Fonseca et al., 2010), Rhinitis Control Assessment Test (Nathan et al., 2010) or visual analogue scales (Bousquet, Combescure, Klossek, Daurès, & Bousquet, 2009) scores).

Despite of allergic rhinoconjunctivitis being a very common phenotype of AR (Papadopoulos et al., 2015), allergic conjunctivitis is frequently considered as a comorbidity rather than an independent risk factor for AR (Bielory, 2010). According to the World Allergy Organization, more than a billion people suffer from allergic conjunctivitis (Pawankar, Canonica, Stephen, Richard, & Michael, 2013). Ocular symptoms are common in patients with AR (50-90%) and contribute to the burden of rhinitis (Rosario & Bielory, 2011), even though they are often underdiagnosed and consequently undertreated. Conjunctivitis is now increasingly recognized as a distinct disorder that may be coupled with more severe disease (Cibella et al., 2015; Garcia-Aymerich et al., 2015).

The second most common rhinitis phenotype is NAR, which is defined as a form of non-infectious rhinitis in which an allergic component it is not identifiable (De Greve et al., 2017; Papadopoulos et al., 2015). NAR can be further subdivided in many phenotypes, such as environmental (occupational, smoking), inflammation (NAR with eosinophilia syndrome or local allergic rhinitis, hormones (pregnancy), drug-induced (rhinitis medicamentosa, non-steroidal anti-inflammatory drugs, aspirin), age (rhinitis of

the elderly), and/or idiopathic rhinitis (none of these triggers are present) (Hellings et al., 2017; Muraro et al., 2016).

Recently, a novel phenotype of rhinitis has been recognized, known as Nasal hyper-reactivity, where the nasal symptoms are induced by exposure to a variety of environmental and endogenous triggers like temperature/humidity changes, physical exercise, stress, and/or exposures to irritants, irrespective of the presence of allergy (Van Gerven, Steelant, & Hellings, 2018). Also, concomitant chronic rhinosinusitis with/without nasal polyps must be considered as anatomical factors that might worsen rhinitis (Hastan et al., 2011).

To help guide appropriate treatment decisions, clarifying the various phenotypes specific to each airway disease is of the highest importance (Heaney & McGarvey, 2017; Kraft, 2011). To this end, biomarkers represent the pillar of stratified medicine allowing for an adequate classification of individuals into subpopulations that differ in their susceptibility to a specific disease or in their response to a singular treatment (Arron, Townsend, Keir, Yaspan, & Chan, 2015; Wagner, 2002). However, the characterization of rhinitis phenotypes is made difficult by the scarcity of distinct biomarkers, even for allergic rhinitis, for which the immunopathogenesis is more clearly defined (Hellings et al., 2017).

COPD

COPD is “a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases” (Vogelmeier et al., 2017). According to WHO estimations, about 65 million people suffer from COPD (Forum of International Respiratory Societies, 2017), with an estimated global prevalence of 11.7% (8.4%–15.0%) in 2010 (Adeloye et al., 2015). In Portugal, it is estimated that about 800.000 subjects having more than 40 years are affected with this condition, with a prevalence of 14.2% (Bárbara et al., 2013).

Over the last two decades, phenotypes of COPD moved away from a forced expiratory volume in 1 second (FEV_1)-centric view to one which also considers clinically relevant domains of the disease such as the level of current symptoms and the history of previous exacerbations (Agusti & MacNee, 2013; Vogelmeier et al., 2017). Characterization of these phenotypes is currently based on clinical manifestations (breathlessness, cough, sputum production, wheezing and chest tightness), patients

characteristics (age, sex, ethnicity), environmental exposures (active/passive smoking, indoor cooking, workplace exposures, ambient pollution, and infections), lung function abnormalities (airflow limitation, bronchial hyperresponsiveness, gas trapping and increased lung volumes, low diffusion capacity, and abnormal pulmonary gas exchange), imagological characterization (lung structural abnormalities, including emphysema, airway wall thickening, and/or bronchiectasis), COPD exacerbations ("frequent exacerbator"), extrapulmonary comorbidities (such as metabolic, lung cancer and cardiovascular), genetic factors (α1-antitrypsin deficiency), and/or systemic inflammation (blood eosinophils count, leukocytes, C-reactive protein, IL-6 IL-8, and fibrinogen) (Brightling et al., 2000; Celli & Agustí, 2018; Han et al., 2010; Singh, Kolsum, Brightling, Locantore, & Agusti, 2014). The serum level of α1-antitrypsin is a well-established biomarker and may be treatable, in α1-antitrypsin deficiency (Chapman et al., 2015). Other biomarkers that have been proposed, despite of none of them has been accepted for COPD could also be useful in future decision-making (Agustí et al., 2012; Kostikas, Bakakos, Papiris, Stolz, & Celli, 2013).

Asthma

Asthma is currently defined by the Global Initiative Program for Asthma as "a heterogeneous disease, usually characterized by chronic airway inflammation; it is defined by the history of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable expiratory airflow limitation" (Gibson, Loddenkemper, Lundbäck, & Sibille, 2013). In Europe, asthma affects about 30 million children and young adults, with a prevalence in northern and western countries among adults aged 18-44 ranging from 3% to 9%. In Portugal, asthma affects 6.8% of the population (Sá-Sousa et al., 2012).

Asthma phenotypes were initially characterized by Rackemann (1927) based on a single dimension and focused on simple classifications. That work described two clinical asthma phenotypes: extrinsic asthma (believed to be due to allergens from outside the body and associated with younger age of onset, environmental triggers, atopy and the presence of other allergic diseases) and intrinsic asthma (due to factors intrinsic to the body associated with older age at onset and the absence of atopy (Rackemann, 1927).

The assumption that asthma must be classified only regarding its allergic component (allergic vs non-allergic asthma) has been increasingly challenged and disproved (The

Lancet, 2006). Asthma is nowadays recognised as a heterogeneous condition, an "umbrella term" for several distinct diseases, caused by a distinct underlying pathophysiological mechanism, having common symptoms but different aetiology, pathogenesis and responses to treatment (Wenzel, 2006). Therefore, under such umbrella term, different phenotypes of asthma may occur.

Over the years, many more different clinical phenotypes of asthma have been described in the literature. Currently, the approaches to phenotyping asthma most commonly used include demographic and risk factors (such as age, obesity, smokers), clinical (such as early/late-onset, symptoms, exercise-induced, health status), response to treatment (such as inhaled and/or oral steroid sensitivity), airway obstruction (variable or partially fixed), type of airway inflammation (eosinophilic, neutrophilic, mixed, pauci-granulocytic), trigger-related, bronchial hyperresponsiveness, exacerbations, and atopy (Agache, Akdis, Jutel, & Virchow, 2012; Hüls et al., 2019; Pavord et al., 2018). Other less frequent phenotypes are based on prognostic factors (brittle asthma, near-fatal asthma and benign asthma) and extensive remodelling (such as thicker small airways, goblet cell hyperplasia and mucus production) (Agache et al., 2012). Moreover, phenotypes of severe asthma arose from several studies from specialized centres (Lefaudeux et al., 2017; Moore et al., 2010; Schatz et al., 2014) and were mainly organized by clinical, trigger or inflammatory characteristics (Chanez et al., 2007; Moore et al., 2010).

Several biomarkers are already available for asthma, particularly severe, and difficult-to-treat asthma (Schleich, Demarche, & Louis, 2016; Wadsworth, Sin, & Dorscheid, 2011), and, in some cases, their use has been incorporated into treatment guidelines (Chung et al., 2014; Global Initiative for Asthma, 2018). The ideal asthma biomarker links the disease pathogenetic mechanism (endotype) with the phenotype. Also, biomarkers should predict disease exacerbation, severity, response to treatment, stability over time, and they must be easily replicated across populations with different genetic backgrounds (Agache et al., 2012; Agache, 2013). However, biomarkers substantially vary across age, severity, complexity and in time (Agache, 2014).

The most scrutinized biomarkers in asthma are related to type 2 immune response associated with atopy and eosinophilic inflammation (Robinson et al., 2017; Woodruff et al., 2009), however there are also clinically defined variant forms of the disorder which are independent of atopy (Macfarlane et al., 2000). Asthma is usually driven by T cell activation, where T helper (Th) 2 cytokines IL-4, IL-5, IL-9 and IL-13 are thought to play a role in the pathophysiology of allergic and non-allergic asthma (Lemanske &

Busse, 2003; Wills-Karp, 1999). On the other hand, it has been shown that asthma, as a heterogeneous disease, involves other mechanisms that are not so well understood and respond poorly to corticosteroid therapy (non-type 2-mediated) (Hastie et al., 2013; Woodruff et al., 2009). A striking example is the interferon-γ, a classical Th1 cytokine, that may cause severe airway inflammation (Hansen, Berry, DeKruyff, & Umetsu, 1999). Subsequently, the phenotypic description of patients with asthma according to the type of underlying inflammation used to distinguish between the so-called "type 2-high" and "type 2-low" phenotypes (depending on the evidence or not of Th2 typical biomarkers assessed in each single patient) (Woodruff et al., 2009).

There has been a recent rise in the number of studies trying to identify asthma phenotypes based on non-invasive, reliable, and easy-to-assess biomarkers, such as blood eosinophils (B-Eos) count, fraction of exhaled nitric oxide (FeNO), serum total IgE, sputum eosinophils and neutrophils, C reactive protein (CRP), eosinophil cationic protein (ECP), serum amyloid A, and/or periostin (Haldar et al., 2008; Jia et al., 2012; Malinovschi, Fonseca, Jacinto, Alving, & Janson, 2013; Rufo, Taborda-Barata, & Lourenço, 2013; Silkoff et al., 2017). In addition, promising novel biomarkers were discovered in the field of breath metabolomics (e.g. volatile organic compounds and exhaled breath condensate) (Donnelly, 2010; Ibrahim et al., 2011). However, several of these biomarkers are at a research level and currently unavailable for clinical practice.

Although induced sputum allows a direct insight of airway inflammation, having the highest diagnostic value for asthma (Henry, 2002), it is not widely used, mainly because it requires extremely invasive and complicated techniques, thus becoming hard to use as a routine diagnostic tool (Pavord et al., 2002). Instead, blood eosinophils count is obtained as proxy of airway eosinophilia (Wagener et al., 2015).

There are different cut-offs of B-Eos counts when assessing different populations: 150/mm^3, 300/mm^3 (most used), 400/mm^3 and 450/mm^3 (less used). The variation for this cut-off is so great, even though Schleich et al. (2014) identified an ideal cut-off of 188/mm^3 for blood eosinophils to detect sputum eosinophilia of ≥3% in a population with severe asthma with a sensitivity of 72% and specificity of 73%. Despite not being very sensitive and specific for predicting a sputum eosinophil of ≥3%, B-Eos cut-offs may help the stratification of type 2-high versus type 2-low asthma phenotypes (Froidure et al., 2016; Katial et al., 2017). Recently, the ELEN index, a model that combines different variables obtained by a complete blood count (CBC), estimated a high or low probability that the patient has eosinophilic airway disease (Khatry et al., 2015). However, further validation is required.

Other biomarkers are obtained in exhaled air. FeNO is a non-invasive biomarker that primarily signals airway inflammation triggered by IL-4 and IL-13 (Alving & Malinovschi, 2010), that is both easy and quick to measure. The use of FeNO as a diagnostic and decision-support tool for asthma management has been gradually increasing in routine care (Dweik et al., 2011). The American Thoracic Society (ATS) guidelines recommended the following cut-offs: if FeNO <25 parts per billion (ppb) (20 ppb in children <12 years old) there is a low likelihood of eosinophilic inflammation and corticosteroid responsiveness; if FeNO >50 ppb (>35 ppb in children) there is a high probability of eosinophilic airway inflammation. The intermediate FeNO range of 25-50 ppb (20-35 ppb in children) should be interpreted with consideration of the clinical context (Dweik et al., 2011). However, the guidelines recognize that these fixed cut-offs have low quality of evidence. Recently, a novel model for prediction of reference FeNO values has been proposed (Jacinto et al., 2018). This could be a useful approach to the interpretation of FeNO in clinical practice, however it should be further validated in large samples.

Both FeNO and B-Eos count partially reflect different inflammatory pathways, representing a local (associated more closely with IL-5-driven) and a systemic type 2-marker (mostly dependent on IL-4/IL-13-driven), respectively (Silkoff et al., 2017). Moreover, there appears to be an additive role of biomarkers, such as B-Eos and FeNO, in relation to recent asthma morbidity (Malinovschi et al., 2013). These two biomarkers are not interchangeable and the use of both biomarkers in combination may allow for better targeted and personalized treatment for at least certain subsets of asthma patients (Malinovschi et al., 2013; Malinovschi, Janson, Borres, & Alving, 2016). However, this combination of B-Eos and FeNO to better provide information on the site of inflammation, as has been shown in asthma (Malinovschi et al., 2013), may not be possible in patients with COPD, possibly reflecting the variable effect of smoking and airway infection on FeNO in patients with more complex airway disease (Schleich, Corhay, & Louis, 2016).

Other distinct subgroups of asthma phenotypes are increasingly being reported due to its specific characteristics, such as steroid therapy resistance and lack of inflammatory markers: e.g. subjects with asthma without evidence of type 2-inflammation; obese asthmatic subjects; and patients with asthma-COPD overlap syndrome (ACOS) (Carr, Zeki, & Kraft, 2018; Christenson et al., 2015; Gibeon et al., 2013; Kobayashi, Hanagama, Yamanda, Ishida, & Yanai, 2016). Therefore, there is an increase need for improving the definition of asthma phenotypes.

1.4.2 Data-driven asthma phenotypes

Asthma phenotypes are still classified based on single dimensions of the disease, such as clinical symptoms, triggers, pathology or patterns of airway obstruction (Borish, 2016; Hekking & Bel, 2014; Hirano & Matsunaga, 2018; Wenzel, 2012), failing to account for its complexity and heterogeneity. In an attempt to explore the pathophysiology of specific asthma subgroups and help stratify patients for targeted therapies, unsupervised (data-driven) approaches are being applied in airways disease to identify “novel” accurate and distinct phenotypes, taking into account the multidimensional characteristics of the disease (Haldar et al., 2008; Moore et al., 2010; Siroux et al., 2011; Wu et al., 2014).

There is a clear heterogeneity regarding asthma phenotypes using unsupervised statistical methods since the initial work of Haldar et al. (2008) that sparked further interest in clustering methodology. These studies derived from different samples, with different subjects’ characteristics, multiple variables/dimensions chosen to be analysed and clustering methods (e.g., k-means or LCA). Despite, there have been studies that identified clusters mainly coincident with other larger-scale cluster analysis (Loureiro, Sa-Couto, Todo-Bom & Bousquet, 2015; Loza et al., 2016; Wu et al., 2014), there is lack of statistical confirmation of the differences found in inter- and intra-clusters. Therefore, a systematic review of the adult asthma phenotypes derived with data-driven methods, using variables easily measurable in a clinical setting, is being undertaken by our research team, as described in Appendix I.

All the 52 studies included were published in the last 11 years and recruited patients mostly from specialized centres (n=41; 79%). The most frequent number of phenotypes identified per study was 4 and 5 phenotypes. The studies identified from our literature search which used distance-based approaches for subtyping asthma are shown in Table 1 and those with model-based approaches in Table 2.

Table 1. Studies that identified asthma phenotypes using distance-based approaches, stratified by the data-driven method applied.

Study ID (Author Year)	**Label**
Hierarchical cluster analysis	
Baptist 2018	- “Late-onset asthma” - “Mildest asthma” - “Atopic, long duration of asthma” - “The most severe asthma”
Delgado-Eckert 2018	- “Mild-to-moderate” - “Severe asthma”
Fingleton 2017	- "Severe late-onset asthma/COPD overlap group" - "Moderately severe early-onset asthma/COPD overlap group" - "Moderate to severe asthma group with type 2 predominant disease" - “Early-onset, minimal airflow obstruction” - “Late-onset, minimal airflow obstruction”
Fingleton 2015	- “Moderate-to-severe childhood-onset atopic asthma” - “Asthma-COPD overlap” - “Obese-comorbid” - “Mild childhood-onset atopic asthma” - “Mild intermittent”
Khusial 2017	- “Early atopic” - “Late onset female” - “Reversible” - “Smokers” - “Exacerbators”
Konno 2018	- “Early-Onset, Atopic, Mild Eosinophilic” - “Late-Onset, Low T-Helper Cell Type 2-relatedIndices” - “Late-Onset, Fixed Airflow Limitation, intense T-helper cell type 2-related indices”

	- “Late-Onset, Fixed Airflow Limitation, low T-helper cell type 2-related indices” - “Female Predominance, High BMI, and Intense T-helper Cell Type2–related Indices”
Moore 2010	- “Early onset atopic asthma with normal lung function treated with two or fewer controller medications and minimal health care utilization” - “Early-onset atopic asthma and preserved lung function but increased medication and health care utilization” - “Older obese women with late-onset nonatopic asthma, moderate reductions in FEV_1, and frequent oral corticosteroid use” - “Severe airflow obstruction with bronchodilator responsiveness, childhood onset and atopic” - “Female, later-onset disease and less atopy, with severe airflow obstruction with BD responsiveness”
Qiu 2018	- “Female, small degree of airflow obstruction and early-onset, neutrophilic and mixed granulocytic inflammation” - “Eosinophilic inflammation, severe airflow obstruction” - “Female, neutrophilic with mixed granulocytic asthma, moderate degree of reduction in FEV_1” - “Eosinophilic with mixed granulocytic inflammation, severe airflow obstruction”
Schatz 2014	- “White female, adult onset, without aspirin sensitivity, lower total IgE levels” - “Atopic dermatitis” - “Male sex” - “Nonwhite race” - “Aspirin sensitivity”
Seino 2018	- “Elderly, severe, poorly controlled asthma, possible adherence barriers” - “Elderly with a low BMI and no significant adherence barriers but had severe, poorly controlled asthma” - “Younger, with a high BMI, no significant adherence barriers, well-controlled asthma, no severely affected”
SendÝn-Hernßndez 2018	- “Intermittent or mild persistent asthma, without family antecedents of atopy, asthma, or rhinitis, lowest total IgE levels” - “Mild asthma with a family history of atopy, asthma, or rhinitis, and intermediate levels of total IgE” - “Moderate-severe persistent asthma with corticosteroids and long-acting b-agonists, with high total IgE levels”
Sutherland 2012	- “Nonobese female asthmatics” - “Nonobese male asthmatics” - “Obese uncontrolled asthma” - “Obese well-controlled asthma”

Weatherall 2009	- “Severe and markedly variable airflow obstruction with features of atopic asthma, chronic bronchitis and emphysema” - “Features of emphysema alone” - “Atopic asthma with eosinophilic airways inflammation” - “Mild airflow obstruction without other dominant phenotypic features” - “Chronic bronchitis in nonsmokers”
Ye 2017	- “Early onset atopic asthma” - “Small airway obstruction and atopic asthma” - “Late-onset and non-atopic asthma” - “Severe airflow obstruction and obvious airway remodeling”
Youroukova 2017	- “Late-onset, non-atopic asthma with impaired lung function” - “Late-onset, atopic asthma” - “Late-onset, aspirin sensitivity, eosinophilic asthma” - “Early-onset, atopic asthma”
K-means cluster analysis	
Amelink 2013	- “Severe eosinophilic inflammation, persistent airflow limitation” - “Obese female symptomatic, high health care utilization” - “Mild-to-moderate, well-controlled asthma with normal lung function”
Deccache 2018	- "Rather confident" - "Rather committed" - "Rather questing" - "Rather concerned"
Jang 2013	- “Well-preserved pulmonary function” - “Female, severe airway obstruction” - “Female, bronchial hyperresponsiveness” - “Male, smokers”
Kim 2017	- “Early-onset atopic asthma with preserved lung function” - “Late-onset non-atopic asthma with impaired lung function” - “Early-onset atopic asthma with severely impaired lung function”

	- “Late-onset non-atopic asthma with well-preserved lung function”
Konstantellou 2015	- “Not related to persistent airflow obstruction, non-atopic patients, without high-dose ICS or OCS” - “Persistent airflow obstruction, atopic, with high dose ICS and OCS” - “Not related to persistent airflow obstruction, atopic, without high-dose ICS or OCS”
Lee 2017	- “Near-normal” - “Asthmatic” - “COPD” - “Asthmatic-overlap” - “COPD-overlap”
Musk 2011	- “Normal males” - “Normal females” - “Atopic younger” - “Obese females” - “Atopic with high eNO” - “Atopic males with poor FEV_1” - “Atopic with BHR”
Park 2013	- “High asthma control test (ACT) scores, low FEV_1, high systemic corticosteroid use” - “Lowest FEV_1, ACT, and quality of life questionnaire for adult Korean asthmatics (QLQAKA) scores, high systemic corticosteroid use” - “Low $FEV_1$and systemic corticosteroid use, improvement in subjective symptoms over time” - “High FEV_1, the lowest systemic corticosteroid use, and had high ACT and QLQAKA scores”
Park 2015	- “Long symptom duration and marked airway obstruction” - “Female dominance and normal lung function” - “Smoking male dominance and reduced lung function” - “High body mass index and borderline lung function”
Rootmensen 2016	- “History of extensive cigarette smoking, airway obstruction without signs of emphysema” - “Features of the emphysematous type of COPD” - “Characteristics of allergic asthma” - “Features suggesting an overlap syndrome of atopic asthma and COPD”

Tanaka 2018	- “Rapid exacerbation, young to middle-aged, hypersensitive to environmental triggers and furred pets” - “Fairly rapid exacerbation, middle-aged and older and low perception of dyspnea” - “Slow exacerbation, high perception of dyspnea, smokers, and chronic daily mild-moderate symptoms”
Wu 2018	- “Atopic nasal polyps and comorbid asthma (NPcA)” - “Smoking NPcA” - “Older NPcA”
Zaihra 2016	- “Severe asthmatics and predominantly late-onset disease” - “Female, severe asthmatics, with higher BMI” - “Severe asthma with reductions in pulmonary function at baseline, early onset, atopic” - “Moderate asthmatics and the majority had good lung function”
Two-step cluster analysis	
Haldar 2008	- “Early-onset atopic asthma” - “Obese female with no eosinophilic inflammation” - “Benign asthma” - “Early-onset, symptom-predominant group with minimal eosinophils” - “Late-onset, male predominant, eosinophilic inflammation with few symptoms”
Hsiao 2018	Female: - “Late-onset, normal BMI, nonatopy, low neutrophils, low eosinophils, normal lung function” - “Young adults with atopy, normal BMI, high blood eosinophils, low neutrophils” - “Late-onset, obesity, high neutrophils, low eosinophils and IgE” Male: - “Late-onset, with low IgE and blood eosinophils, normal BMI, normal lung function” - “Young adults with atopy, current smoking, and high blood neutrophils” - “Late-onset, ex-smokers, high blood eosinophils”
Ilmarinen 2017	- “Nonrhinitic asthma” - “Smoking asthma” - “Female asthma” - “Obesity-related asthma” - “Early-onset atopic adult asthma”

Kim 2013	- “Smoking asthma” - “Severe obstructive asthma” - “Early-onset atopic asthma” - “Late-onset mild asthma”
Lemiere 2014	- “Exposition to high molecular-weight (MW), normal lung function, no taking ICS, atopy” - “Exposition to high MW, lower lung function, taking ICS, atopy” - “Only exposed to low MW agents, lower lung function, taking ICS less atopy”
Newby 2014	- “Early-onset atopic” - “Late-onset obese” - “Least severe disease” - “Late-onset eosinophilic” - “Significant severe airflow obstruction”
Serrano-Pariente 2015	- “Elderly with clinical and therapeutic criteria of severe asthma” - “High proportion of respiratory arrest, impaired consciousness level and mechanical ventilation” - “Younger, insufficient anti-inflammatory treatment and sensitization to *Alternaria alternata* and soybean”
Wang 2017	- “Allergic asthma” - “Fixed airflow limitation” - “Low socioeconomic status” - “Female, mild asthma with slight airway obstruction and low exacerbation risk” - “Male, mild asthma with slight airway obstruction and low exacerbation risk”
Wu 2014	- “Severe asthma, less asthma symptoms and better AQLQ scores” - “Early-onset allergic asthma with low lung function and eosinophilic inflammation” - “Later-onset, mostly severe asthma with nasal polyps and eosinophilia” - “Persistent inflammation in blood and bronchoalveolar lavage fluid and exacerbation” - “Severe asthma, Hispanic woman, frequent symptoms, low AQLQ scores, high degree of allergic sensitization” - “Severe asthma, normal lung function and no symptoms”
K-medoids cluster analysis	
Lefaudeux 2017	- “Severe late-onset asthma with airway obstruction, high BMI, smoking, and OCS use” - “Severe asthma with airway obstruction and OCS use but no smoking history”

	- “Severe asthma with female predominance, high BMI, frequent exacerbations, and OCS use but no history of smoking or airway obstruction” - “Moderate to severe” - “Well controlled”
Loureiro 2015	- “Early onset mild allergic asthma” - “Moderate allergic asthma, with long evolution, female prevalence and mixed inflammation” - “Allergic brittle asthma in young females with early disease onset and no evidence of inflammation” - “Severe asthma in obese females with late disease onset, highly symptomatic despite low Th2 inflammation” - “Severe asthma with chronic airflow obstruction, late disease onset and eosinophilic inflammation”
Loza 2016	- "Mild, good lung function, early onset" - "Moderate, hyper-responsive, eosinophilic" - "Mixed severity, predominantly fixed obstructive, non-eosinophilic and neutrophilic" - "Severe uncontrolled, severe reversible obstruction, mixed granulocytic"
Sekiya 2016	- “Younger-onset asthma with severe baseline symptom” - “Female-predominant elderly asthma” - “Allergic asthma without baseline ICS treatment” - “Male-predominant COPD-overlapped elderly asthma” - “Asthma with almost no baseline symptoms”
Hierarchical clustering followed by K-means and K-medoids	
Brinkman 2018	- “Middlemost cluster” - “Neutrophilic inflammation predominant” - “Eosinophilc inflammation predominant”

Legend: HC: Hierarchical clustering; PAM: Partition-around-medoids

Table 2. Studies that identified asthma phenotypes using model-based approaches, stratified by the data-driven method applied.

Study ID (Author Year)	**Label**
Latent class analysis	
Couto 2015	- “Atopic asthma” - “Sports asthma”
Jeong 2017	- “Persistent multiple symptom-presenting asthma” - “Symptom-presenting asthma” - “Symptom-free atopic asthma” - “Symptom-free non-atopic asthma”
Makikyro 2017	Female: - “Controlled, mild asthma” - “Partly controlled, moderate asthma” - “Uncontrolled asthma, unknown severity” - “Uncontrolled, severe asthma” Male: - “Controlled, mild asthma” - “Poorly controlled asthma, unknown severity” - “Partly controlled, severe asthma”
Siroux 2011	- ‘‘Active treated allergic childhood-onset asthma’’ - ‘‘Active treated adult-onset asthma’’ - ‘‘Inactive/mild untreated allergic asthma’’ - ‘‘Inactive/mild untreated nonallergic asthma’’ - ‘‘Active treated allergic childhood-onset asthma’ - ‘‘Active treated adult-onset asthma’’ - ‘‘Inactive/mild untreated allergic childhood-onset asthma’’ - ‘‘Inactive/mild untreated adult onset asthma’’

van der Molen 2018	- “Confident and self-managing" - "Confident and accepting of their asthma" - "Confident but dependent on others" - "Concerned but confident in their health care professional (HCP)" - “Not confident in themselves or their HCP”
Factor analysis	
Alves 2008	- “Treatment-resistant, more nocturnal symptoms and exacerbations” - “Persistent airflow limitation” - “Allergic rhinosinusitis, nonsmokers, reversible airflow obstruction” - “Aspirin intolerance”
Moore 2014	- “Mild-to-moderate early-onset allergic asthma with paucigranulocytic or eosinophilic sputum inflammatory cell patterns” - “Mild-to-moderate early-onset allergic asthma with paucigranulocytic or eosinophilic sputum inflammatory cell patterns, OCS use” - “Moderate-to-severe asthma with frequent health care use despite treatment with high doses of inhaled or oral corticosteroids, normal lung function” - “Moderate-to-severe asthma with frequent health care use despite treatment with high doses of inhaled or oral corticosteroids, reduced lung function”
Latent transition analysis // Expectation-maximization	
Boudier 2013	- “Allergic, few symptoms, no treatment,” - “Non-allergic, few symptoms, no treatment” - “Non-allergic, high symptoms, treatment” - “Allergic, high symptoms, treatment, BHR” - “Allergic, moderate symptoms, BHR” and - “Allergic, moderate symptoms, normal lung function” - “Non-allergic, moderate symptoms, no treatment”
Janssens 2012	- “Well-controlled asthma” - “Intermediate asthma control” - “Poorly controlled asthma”

Legend: HCP: Health care professional; OCS: oral corticosteroids; BHR: Bronchial hyperresponsiveness

The distance-based approaches were the most applied (n=43; 83%), particularly the following data-driven methods: hierarchical methods, k-means and two-step cluster analysis (Table 1). Of the 9 studies that applied model-based approaches (Table 2), the most method used was LCA (n=5).

The most frequent data-driven phenotypes were the allergic and non-allergic asthma with multiple variants. Also, phenotypes stratified by the predominant sex and age of asthma onset was frequently observed in both distance- and model-based approaches.

Although the two approaches proved capable of recovering cluster structure with clinical meaning, they are not interchangeable and must be applied regarding the respective specifications of each statistical method.

Variables were matched and categorized into six groups: clinical, functional, socio-demographic, inflammation, atopy, and other (psychological/behaviour variables, asthma-related medication use and healthcare use).

Figure 4 presents the proportions of studies that used each dimension for asthma phenotyping. The most used dimensions were variables regarding clinical and functional data, however, a substantial difference was observed according to the sample characteristics, data availability, study aim and data sources.

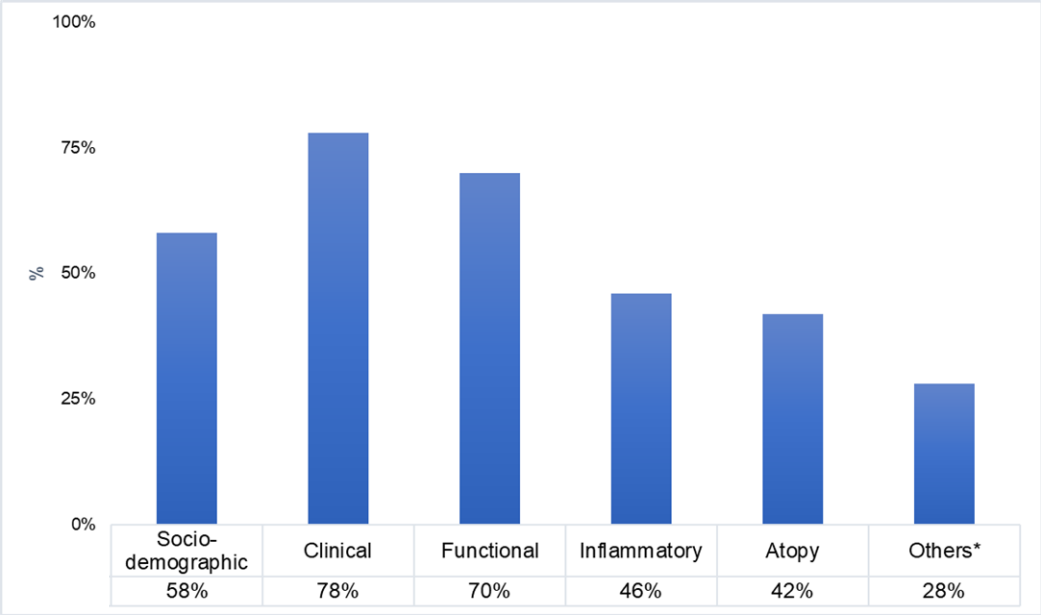


Figure 4. Proportion of each dimension chosen for the data-driven phenotyping.
*Psychological/behavioural variables, asthma-related medication use, healthcare use.

Moreover, studies evaluating the consistency of phenotypes (n=15; 29%) followed at least one of the criteria: longitudinal stability, cluster repeatability, reproducibility, internal and external validity. However, studies with population-based samples and reporting longitudinal consistency of data-driven phenotypes are scarce, and generalization to the “overall” asthma population may be limited.

As expected, a high degree of variability regarding the data-driven methods and variables applied in the models was observed. Therefore, this systematic review intended to provide important information on limitations of data-driven phenotype studies in asthma and the need for improvement in future studies of population-based studies.

2. Aims

To answer the question: “Can the phenotypes of chronic diseases of the airways be better classified/refined using multidimensional models and data-driven approaches?”, we aimed:

- To explore multidimensional models, supported by advanced data-driven statistical methods, for classification of phenotypes of chronic diseases of the airways, based on clinical, functional, and immuno-inflammatory characteristics.

As specific aims, we intended:

1- To describe the proportion of overlap of five theory-driven asthma phenotypes of adults from the US National Health and Nutrition Examination Survey (NHANES) population-based study, and to examine their association with asthma-related outcomes;

2- To compare previously defined theory-driven asthma phenotypes with newly data-driven ones, identified by latent class analysis (LCA), in the US adults with current asthma from the NHANES.

3- To identify distinct phenotypes of allergic respiratory diseases, using LCA, in adults from the Portuguese Control and Burden of Asthma and Rhinitis (ICAR) nationwide study and to and to explore the most relevant clinical variables that could be used to distinguish each class.

3. Studies

3.1 Study I

Amaral, R., Fonseca, J. A., Jacinto, T., Pereira, A. M., Malinovschi, A., Janson, C., & Alving, K.

"Having concomitant asthma phenotypes is common and independently relates to poor lung function in NHANES 2007–2012".

Clinical and Translational Allergy. 2018; 8:13.

The aims of study I (Appendix II) were to describe the proportion of overlap of five commonly reported asthma phenotypes: asthma with obesity (AwObesity), asthma with concurrent COPD (AwCOPD), B-Eos-high, FeNO-high and B-Eos&FeNO-low asthma, and to examine the association of their overlap with asthma-related outcomes, using population-based data from the National Health and Nutrition Examination Surveys (NHANES), 2007–2012.

Methods

Study setting

The NHANES is a nationally representative survey of the civilian, non-institutionalized US population that uses a complex stratified, multistage probability sampling. Also, NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey combines interviews and physical examinations. The interview includes demographic, socioeconomic, dietary, and health-related questions, while the examination consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel (CDC, 2017). The National Center for Health Statistics, Ethics Review Board approved NHANES protocol, and all participants gave written informed consent.

Subjects selection

Six survey years (NHANES 2007–2012) were analyzed, resulting in 30,442 individuals of all ages (Figure 5). We included adults (≥ 18 years-old) with current asthma (n=1059), defined by a positive answer to the questions: “Has a doctor ever told you that you have asthma?” together with “Do you still have asthma?”, and either “wheezing/whistling in the chest in the past 12 months” or “asthma attack in the past 12 months.”

Variables

Demographic characteristics, such as age, gender, body mass index (BMI), race/ethnicity, and educational status were analyzed. B-Eos count, FeNO and spirometric measurements, collected at the NHANES Mobile Examination Center were also examined. FeNO and spirometric measurements not fulfilling ATS/ERS recommendations (Miller, 2005; Silkoff, 2005) were excluded (n = 653). After predicted values of basal FEV_1 and FEV_1/functional vital capacity (FEV_1/FVC) were calculated (Hankinson, Odencrantz, & Fedan, 1999), with a correction factor for ethnicity (Hankinson et al., 2010), abnormal lung function was defined if either one of them were less than the lower limit of normal (LLN), defined as lower fifth percentile of the reference population (Pellegrino, 2005).

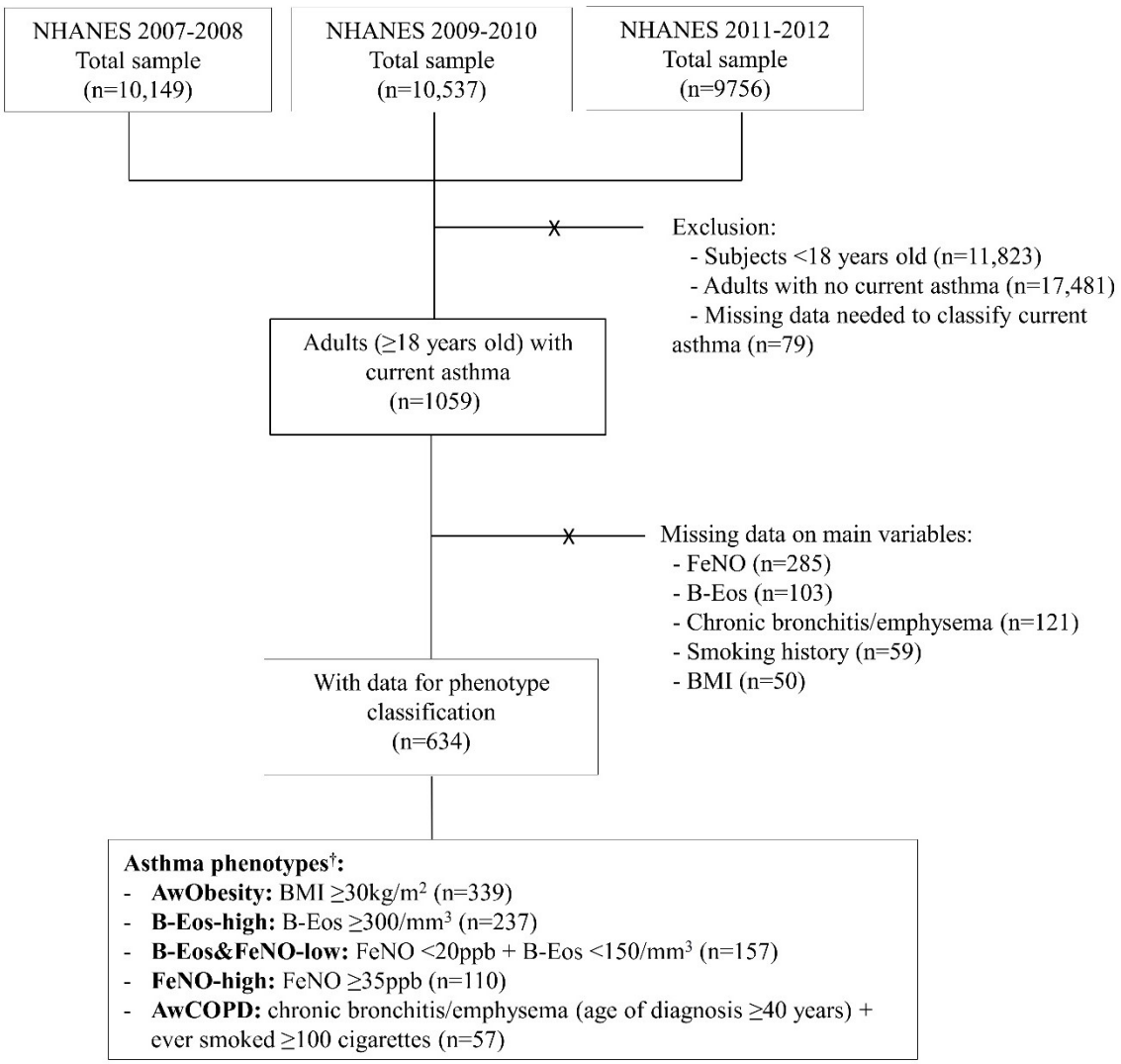


Figure 5. Flowchart of the study analysis. †Seventy-seven patients were considered "non-classified" (non-single and non-multiple phenotype)

B-Eos were part of the complete blood counts and assessed on a Beckman Coulter MAXM® instrument (Beckman Coulter, Fullerton, Calif). FeNO measurements were performed using the analyzer NIOX MINO® (Aerocrine, Solna, Sweden). A detailed description can be found elsewhere (CDC, 2011a, 2011b, 2013).

Demographic characteristics, such as age, gender, BMI, race/ethnicity, and educational status were analyzed. The following categorization was performed:

- *BMI* was calculated and classified based on the definition of the World Health Organization (WHO, 1995): underweight (BMI ≤18.4 kg/m^2); normal (18.5-24.9 kg/m^2), overweight (25–29.9 kg/m^2), and obese (≥30 kg/m^2);

- *Educational status* was divided in: less than high school (<high school) and completion or greater than high school (≥high school).
- *Current smoking* was considered if participants had a positive answer to both questions: “Have you smoked ≥100 cigarettes during lifetime” and “Do you now smoke cigarettes?”. If participants answered positively to the first question but negatively to the second one, they were considered ex-smokers.

Moreover, clinical variables were defined as:

- *Self-reported asthma attacks* and *asthma-related emergency department (ED) visits* (Yes/No), in the past 12 months were analyzed.
- *Work/school absenteeism* was defined as having at least one day lost at work/school due to wheezing (Yes/No).
- *Asthma symptoms* were evaluated with the following questions regarding the last 12 months: “Had wheezing/whistling in your chest?”; “Had disturbed sleep due to wheezing?”; “Had dry hard cough at night, not associated with a cold for at least 14 days in a row?”; “Had wheezing during/after exercise?”; and “Had limited activity due to wheezing?”.
- *Self-reported rhinitis* was defined by an affirmative answer to “During the past 12 months, have you had an episode of hay fever?”.
- *Use of reliever/rescue medication* for asthma was considered if the participant used short-acting β_2-agonist, anticholinergic, or inhaled corticosteroids (ICS)/formoterol.
- *Controller medication* included: ICS; leukotriene modifiers; long-acting inhaled β_2-agonist (LABA) and ICS combination; ICS/LABA associated with methylxanthine, cromoglycate, and/or oral corticosteroids. Details on the prescription medication data collection in NHANES 1999–2012 are available elsewhere (CDC, 2014).

Asthma phenotypes definition

A B-Eos count ≥ 300/mm^3 was used to define a B-Eos-high asthma phenotype (Castro et al., 2015; Sally Wenzel et al., 2013), while FeNO-high was defined as FeNO ≥ 35 ppb (Dweik et al., 2010). Asthma patients with both B-Eos <150/mm^3 and FeNO <20 ppb were categorized as B-Eos&FeNO-low asthma (McGrath et al., 2012). Additionally, we considered subjects with either B-Eos-high or FeNO-high as having “Type 2-high” asthma. The AwObesity phenotype was defined by a BMI ≥ 30 kg/m2 in individuals with current asthma (Flegal, Carroll, Kuczmarski, & Johnson, 1998). Finally, the AwCOPD phenotype was considered if participants ≥ 40 years-old had concurrent asthma and COPD, defined by a positive answer to “Has a doctor ever told you that you have chronic bronchitis/emphysema”, with age of diagnosis ≥ 40 years and having

self-reported smoking history (being either a current or ex-smoker) (Buist et al., 2007; Rossi et al., 2017).

Statistical analysis

In accordance with the NHANES sampling design, the weights for each full sample 2-year mobile examination center were used to obtain weighted percentages adjusted to the US adult population. Categorical variables were described as frequencies and weighted proportions, and continuous variables were described as median and first and third quartiles (Q1–Q3). Chi square test and Mann–Whitney U-test were used to compare groups. To explore the association of concomitant (having at least 2 concurrent) phenotypes with each asthma-related outcome we performed multivariate logistic regression modelling. Separate models were run using each asthma-related outcome and abnormal lung function as dependent variable and having multiple phenotypes as independent variables. Adjustments were also made for potential confounders: sex, age, race, current smoking and rhinitis. Adjusted odds ratios (aOR) with 95% confidence intervals (95% CI) were presented, and model fit was assessed using the *svylogitgof* function (Archer, Lemeshow, & Hosmer, 2007). According to age (<40 or ≥40 years-old), a four- or five-set Venn-Euler diagram was used to quantify the proportion of individuals with different asthma phenotypes and to illustrate the overlap.

The diagrams were created using R software version 3.2.0 (“VennDiagram”, “venneuler” and “reshape2” packages) and all statistical analyses were performed in Stata version 13.1 (StataCorp, TX, USA), using the survey command to account for the complex sampling design and weights in the NHANES. The *MI* command was used to perform sensitivity analysis by multiple-imputation of missing values; however, to create the Venn-Euler diagrams, a listwise deletion for missing data was applied. A p-value <0.05 was considered statistically significant.

Results

Of the 18,619 adults included in NHANES 2007–2012 datasets, 1059 (5.6% [95% CI 5.1–5.9]) had current asthma (Figure 5). Of these, 63% were female, and the median (Q1–Q3) age was 48.0 (32.0–62.0) years. After excluding subjects with missing data on the main variables, 634 individuals were included for phenotype classification (Figure 5). Despite having all information available, 77 patients did not meet the criteria for any of the defined asthma phenotypes and were considered “non-classified”. These were non-obese subjects with asthma who did not meet the criteria for COPD, had B-Eos values ranging between 150 and 300/mm^3, and FeNO ranging 20–34 ppb.

Phenotypes and overlap description

The weighted proportions of asthma phenotypes were (in descending order): 49% for AwObesity, 36% for B-Eos-high asthma, 26% for B-Eos&FeNO-low asthma, 18% for FeNO-high asthma, and 8% for AwCOPD (Table 3). Demographic and clinical characteristics among all 5 asthma phenotypes and the “non-classified” group are described in Table 4. There is a female predominance among all phenotypes, particularly in the B-Eos&FeNO-low (78%). Subjects with AwCOPD phenotype were the oldest group (median [Q1–Q3]: 61.0 [52.0–69.0] years-old), with the lowest proportion of individuals that had ≥ high school and lowest FEV_1/FVC (0.63 [0.50–0.75]), comparing to the other phenotypes.

Table 3. Characteristics of adults with current asthma: included and excluded from phenotype classification and stratified by single or multiple phenotypes.

Demographic characteristics n (weighted %)	**Included subjects** n=634	**Excluded subjects** n= 425	**p-value***	**Single phenotype†** n= 271	**Multiple phenotypes†** n= 286	**p-value***
Female gender	410 (63)	261 (64)	.93	174 (64)	192 (66)	.68
Age (yrs), median (Q1-Q3)	44.0 (31.0-57.0)	48.9 (33.7-68.0)	**<0.001**	42.0 (30.0-55.0)	47.5 (34.0-60.0)	**0.003**
BMI (kg/m^2), median (Q1-Q3)	30.8 (25.4-35.9)	31.4 (24.4-35.7)	.99	28.7 (24.2-35.0)	33.7 (30.7-39.0)	**<0.001**
Obesity status						
Underweight (≤18.4 kg/m^2)	2 (0.3)	13 (4)	**<0.001**	1 (0.2)	1 (0.5)	.41
Normal (18.5-24.9 kg/m^2)	142 (25)	98 (27)	.44	78 (30)	20 (8)	**<0.001**
Overweigh (25-29.9 kg/m^2)	151 (26)	77 (19)	0.07	80 (32)	38 (15)	**<0.001**
Obese (≥30 kg/m^2)	339 (49)	187 (49)	.94	112 (38)	227 (77)	**<0.001**
Race and/or Ethnicity						
Hispanic	105 (8)	85 (10)	.30	38 (8)	56 (9)	.24
Non-Hispanic white	323 (74)	184 (63)	**0.005**	135 (72)	138 (72)	.93
Non-Hispanic black	167 (14)	117 (18)	**0.03**	76 (14)	81 (16)	.71
Other Race	39 (4)	39 (7)	0.06	22 (6)	11 (3)	.11
Smoking status						
Current smoker	199 (29)	114 (32)	.40	89 (31)	87 (27)	.46
Ex-smoker	163 (29)	98 (23)	.13	56 (20)	89 (36)	**0.003**
Non-smoker	272 (43)	154 (45)	.53	126 (49)	110 (37)	**0.02**
Education						
≥High school	478 (84)	225 (69)	**<0.001**	205 (86)	209 (81)	.13
Asthma-related medication ‡						
Reliever medication**	276 (41)	202 (48)	.16	113 (37)	132 (47)	.10
Oral corticosteroids	33 (8)	29 (3)	**0.001**	12 (3)	15 (4)	.75
Inhaled corticosteroids§	153 (25)	122 (30)	.14	55 (19)	81 (32)	**.01**
Other control medications||	53 (9)	63 (14)	.19	15 (8)	33 (10)	.50
Asthma phenotype						

AwObesity	339 (49)	-	112 (38)	227 (76)	**<0.001**
B-Eos-high	237 (36)	-	61 (22)	176 (62)	**<0.001**
B-Eos&FeNO-low	157 (26)	-	74 (30)	83 (30)	.95
FeNO-high	110 (18)	-	18 (8)	92 (34)	**<0.001**
AwCOPD	57 (8)	-	6 (2)	51 (17)	**<0.001**
Non-classified††	77 (14)	-	-	-	

Legend: *Chi-square test or Mann-Whitney U-test was used. †Seventy-seven subjects included in the "non-classified" group were considered as missing. ‡Prescribed medication taken in the past 30 days. §Alone or in combination with long-acting inhaled β_2-agonist. || Included long-acting inhaled β_2-agonist (without corticosteroids), leukotriene inhibitors, and mast cell stabilizers. **Short-acting β_2-agonist and/or anticholinergic. †† Subjects with non-single & non-multiple asthma phenotype. Data presented as absolute numbers and proportions weighted for the US population. P-values <0.05 are presented in bold. Yrs: years; BMI: Body Mass Index; Q1: first quartile; Q3: third quartile; BMI: body mass index; AwObesity: Asthma with obesity; AwCOPD: Asthma with concurrent COPD.

Table 4. Demographic and clinical characteristics among all 5 phenotypes and in the "Non-classified" group.

Characteristics n (weighted %)	**AwObesity** n=339	**B-Eos-high** n= 237	**B-Eos&FeNO-low** n= 157	**FeNO-high** n= 110	**AwCOPD** n=57	**Non-classified††** n=77
Female gender	366 (67)	196 (55)	138 (78)	78 (52)	71 (58)	44 (56)
Age (yrs), median (Q1-Q3)	48.0 (34.0-59.0)	47.0 (31.0-59.0)	41.0 (27.0-57.0)	45.0 (30.0-54.0)	61.0 (52.0-69.0)	39.0 (28.0-53.0)
BMI (kg/m^2), median (Q1-Q3)	35.4 (32.5-40.4)	30.3 (25.9-36.8)	28.9 (24.2-33.0)	27.6 (24.9-33.4)	30.3 (25.1-35.1)	24.3 (22.8-27.5)
Race and/or Ethnicity						
Hispanic	94 (9)	78 (11)	30 (8)	29 (10)	18 (4)	11 (6)
Non-Hispanic white	230 (65)	175 (73)	79 (69)	65 (75)	80 (81)	50 (86)
Non-Hispanic black	177 (21)	74 (12)	62 (17)	38 (13)	27 (11)	10 (6)
Other Race	25 (5)	21 (4)	13 (6)	6 (2)	10 (4)	6 (2)

Smoking status						
Current smoker	150 (27)	104 (32)	48 (23)	18 (13)	65 (52)	23 (28)
Ex-smoker	134 (27)	100 (32)	35 (25)	42 (39)	70 (48)	18 (30)
Non-smoker	229 (45)	124 (35)	88 (51)	66 (49)	0 (0)	36 (42)
Education						
≥High school	358 (79)	227 (80)	125 (81)	102 (87)	73 (59)	64 (91)
Asthma-related medication ‡						
Reliever medication**	234 (42)	172 (51)	82 (44)	65 (46)	69 (57)	31 (36)
Oral corticosteroids	36 (5)	17 (5)	4 (2)	10 (7)	15 (7)	2 (2)
Inhaled corticosteroids§	144 (27)	103 (32)	36 (20)	36 (33)	60 (47)	17 (19)
Asthma-related outcomes						
Asthma attack	363 (68)	252 (74)	125 (68)	98 (71)	85 (63)	54 (75)
Asthma-related ED	130 (27)	80 (23)	41 (23)	26 (13)	37 (32)	9 (8)
>2 asthma symptoms	311 (66)	199 (65)	92 (55)	80 (57)	90 (74)	42 (59)
Work/school absenteeism	66 (18)	43 (16)	25 (18)	23 (14)	12 (20)	10 (14)
Lung function						
FEV_1% predicted, median (Q1-Q3)	89.0 (75.6-99.2)	84.1 (75.2-95.4)	93.2 (83.8-100.8)	82.7 (75.9-95.4)	74.0 (62.9-90.1)	89.9 (80.5-103.5)
FEV_1/FVC, median (Q1-Q3)	0.77 (0.62-0.82)	0.74 (0.65-0.80)	0.79 (0.72-0.83)	0.72 (0.66-0.79)	0.63 (0.50-0.75)	0.76 (0.69-0.82)

Legend: †† Subjects with non-single & non-multiple asthma phenotype. ‡Prescribed medication taken in the past 30 days. **Short-acting β_2-agonist and/or anticholinergic. §Alone or in combination with long-acting inhaled β_2-agonist. Data presented as absolute numbers and proportions weighted for the US population. Yrs: years; BMI: Body Mass Index; Q1: first quartile; Q3: third quartile; BMI: body mass index; AwObesity: Asthma with obesity; AwCOPD: Asthma with concurrent (COPD); ED: emergency-department; FEV_1: Forced expiratory volume in 1 second; FEV1/FVC: Forced expiratory volume in 1 second and functional vital capacity ratio; LLN: Lower limit of normality.

When categorized by age, < 40 (n = 227) and ≥ 40 years-old (n = 330), the most prevalent phenotypes were AwObesity (42 and 53%, respectively) and B-Eos-high asthma (34 and 37%). The less ones were FeNO-high asthma (18 and 19%) and AwCOPD (19% in the older group) (Figure 6). The areas of intersection in the four- and five-set Venn-Euler diagrams revealed 5 and 12 overlapping categories, and proportions of 17 and 12% of non-classified asthma subjects, respectively.

In both diagrams, a substantial total overlap was observed: 44% in subjects < 40 years-old and 54% in subjects ≥ 40 years-old. About 40% of the individuals in both age groups had two concomitant asthma phenotypes, 4% of the younger group had 3 concomitant phenotypes and 13% of the older group had ≥ 3 (Table 5 and Figure 6). Furthermore, 1% of the older subjects had four concomitant asthma phenotypes: AwObesity, AwCOPD, FeNO-high, and B-Eos-high asthma.

The most prevalent overlaps in both groups (<40 and ≥40 years-old) were AwObesity together with either B-Eos-high (15 and 12%, respectively) or B-Eos&FeNO-low asthma (13 and 11%) (Figure 6). Moreover, the proportions of subjects having AwObesity together with other phenotypes were high: 53% for the B-Eos-high phenotype, 48% for AwCOPD, 45% for the B-Eos&FeNO-low, and 44% for the FeNO-high phenotype. Also, the proportion of individuals having AwCOPD together with the B-Eos-high phenotype was high (36%), whereas the proportions were lower for the B-Eos&FeNO-low and the FeNO-high asthma phenotypes (15 and 10%, respectively) (data not shown).

In this population, only 12 and 15% of asthma subjects (<40 and ≥40 years-old, respectively) with high B-Eos count had a concomitant high FeNO values (Figure 6). Moreover, the two biomarkers were non-congruent across cut-offs. For example, when comparing groups with B-Eos count <150/mm^3 and 150–300/mm^3, the proportion of asthma subjects having low FeNO (<20 ppb), was not significantly different (Table 6).

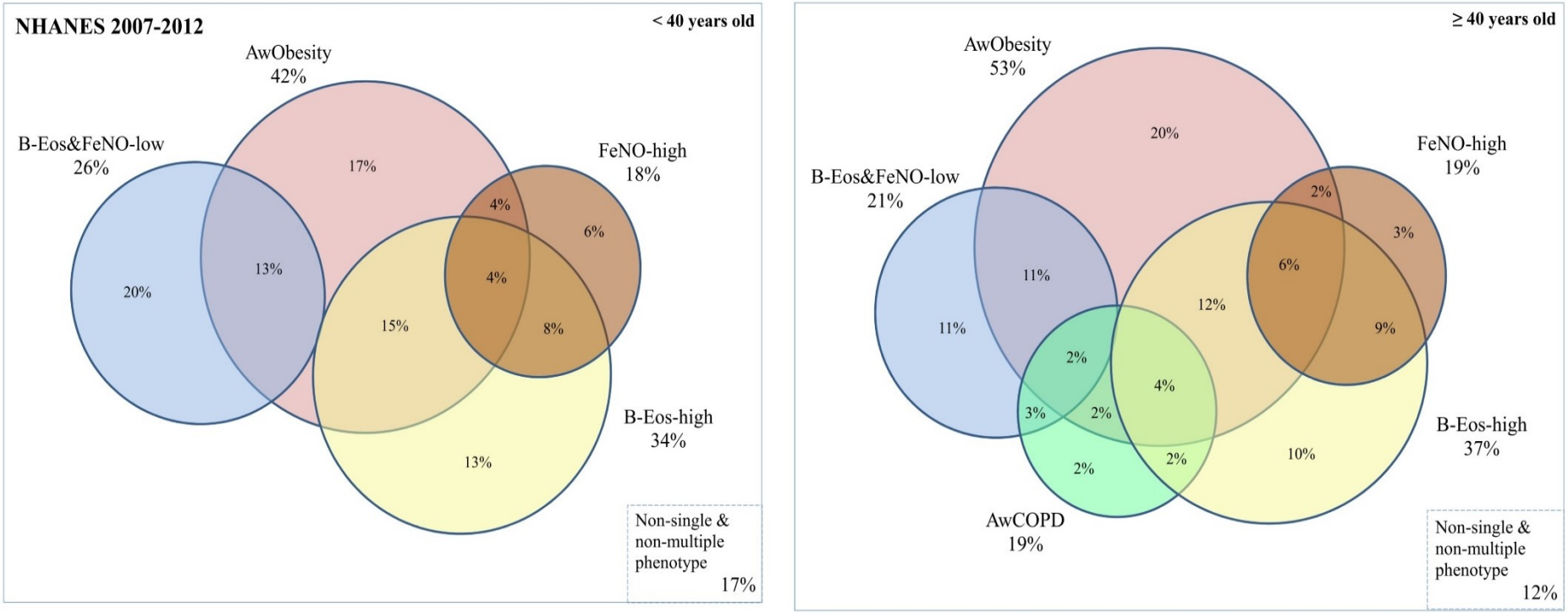


Figure 6. Venn-Euler diagrams quantifying the overlap among the asthma phenotypes, stratified by age. †Seventy-seven patients were considered "non-classified" (non-single and non-multiple phenotype)

Table 5. Distribution and comparisons of the asthma-related outcomes among asthma phenotypes, stratified by age

	Total n (weighted %)	**Asthma attack**	**Asthma-related ED**	**≥2 asthma symptoms**	**Work/school absenteeism**	**Asthma medication**		**Lung function**		
						≥1 reliever medication†	≥2 controller medication	FEV_1 <LLN	FEV_1% predicted §	FEV_1/FVC <LLN
<40 yrs										
1 phenotype	118 (56)	85 (70)	27 (20)	59 (56)	15 (21)	55 (43)	17 (11)	10 (7)	95.6 (90.2-102.1)	26 (20)
2 phenotypes	97 (40)	69 (75)	26 (32)	57 (73)	12 (9)	41 (40)	8 (9)	19 (23)	91.8 (81.3- 99.2)	25 (29)
3 phenotypes	12 (4)	10 (73)	5 (40)	8 (72)	4 (25)	9 (68)	1 (4)	3 (36)	81.9 (75.3- 84.6)	4 (29)
p-value*										
1 vs 2		0.43	0.10	0.052	**0.04**	0.72	0.64	**0.006**	0.07	0.25
2 vs 3		0.92	0.55	0.96	0.07	0.15	0.46	0.43	**0.009**	0.98
1 vs 3		0.86	0.11	0.39	0.73	0.20	0.34	**0.01**	**<0.001**	0.48
≥40 yrs										
1 phenotype	153 (46)	104 (65)	35 (23)	76 (62)	20 (15)	58 (33)	26 (17)	29 (20)	91.2 (81.6-99.2)	34 (26)
2 phenotypes	136 (41)	92 (69)	28 (21)	70 (55)	12 (9)	60 (46)	43 (36)	31 (37)	80.4 (70.0-91.7)	27 (32)
≥3 phenotypes	41 (13)	26 (72)	4 (9)	30 (70)	8 (40)	22 (60)	16 (40)	12 (46)	74.0 (63.0-85.8)	16 (53)
p-value*										
1 vs 2		0.56	0.82	0.38	0.41	0.12	**0.02**	**0.01**	**0.007**	0.46
2 vs ≥3		0.80	0.27	0.30	**0.002**	0.24	0.67	0.47	0.22	0.09
1 vs ≥3		0.50	0.21	0.54	**0.03**	0.053	**0.02**	**0.01**	**0.006**	**0.02**

Legend: *Chi-square test or Mann-Whitney U-test was used. †Short-acting β_2-agonist or/and anticholinergic. §Presented as median (Q1-Q3). Data presented as absolute numbers and proportions weighted for the US population. P-values <0.05 are presented in bold. The 77 subjects included in the "non-classified" group were considered as missing. ED: emergency-department; FEV_1: Forced expiratory volume in 1 second; FEV_1/FVC: Forced expiratory volume in 1 second and functional vital capacity ratio; LLN: Lower limit of normality; Q1: first quartile; Q3: third quartile.

Associations between asthma-related outcomes and phenotype overlap

A comparison of the clinical characteristics of participants with one, two or three or more asthma phenotypes, stratified by age, is presented in Table 4 and no significant differences were observed in any age groups regarding asthma attacks, asthma-related ED, ≥ 2 asthma symptoms, and use of ≥1 reliever. In the older group, the proportion of individuals with work/school absenteeism, ≥ 2 controller medications and with FEV_1/FVC < LLN was significantly higher in participants with concomitant phenotypes than in those with a single phenotype (Table 4). In both age groups, the proportion of patients with FEV_1< LLN was significantly higher when participants presented multiple phenotypes, as well as they presented lower median FEV_1% predicted values.

When analyzing the asthma-related outcomes in subjects with a single phenotype with those having specific combination of asthma phenotypes, the overall findings were that subjects having multiple phenotypes had significantly higher proportion of using ≥ 1 reliever and ≥ 2 controller medications and had decreased lung function, with the exception of those with the B-Eos&FeNO-low phenotype combined with any of the other phenotypes (Table 7 and Table 8).

Moreover, there was a significant association between FeNO and B-Eos categories in both groups, except when comparing the B-Eos groups: <150/mm^3 and 150-300/mm^3 above and below a FeNO cut-off of 20 ppb, in current asthma patients (Table 6).

Table 6. Distribution and comparisons between the FeNO and B-Eos cut-offs used in this study, among individuals with current asthma

		Total n (weigthed %)	**B-Eos** n (weigthed %)			**p-value**		
			Class I <150/mm^3	**Class II** 150-300/mm^3	**Class III** ≥300/mm^3	I vs II	II vs III	I vs III
Current asthma (n= 1,059)			332 (37)	276 (27)	348 (36)			
FeNO	<20ppb	501 (65)	184 (73)	146 (67)	141 (50)	0.12	**<0.001**	**<0.001**
	≥20ppb	273 (35)	60 (27)	66 (33)	130 (50)			
	<35ppb	636 (82)	229 (94)	186 (86)	184 (65)	**0.02**	**<0.001**	**<0.001**
	≥35ppb	138 (18)	15 (6)	26 (14)	87 (35)			

Legend: FeNO: Fraction of exhaled nitric oxide; B-Eos: blood eosinophils.

Table 7. Weighted percentages and comparisons of asthma-related outcomes among subjects with a single asthma phenotype versus: non-classified, and specific combinations of asthma phenotypes

Weighted %	Total	Asthma attack	Asthma-related ED	≥2 asthma symptoms	Work/school absenteeism	Asthma medication		Lung function		
						≥1 reliever medication	≥2 controller medication	FEV_1 <LLN	FEV_1% predicted§	FEV_1/FVC <LLN
Single phenotypes*	43	67	21	59	17	38	14	14	94 (86-97)	23
AwObesity	38	61	27	59	15	31	14	18	92 (84-104)	16
B-Eos-high	22	81	29	70	19	47	15	17	95 (84-99)	27
FeNO-high	8	80	6	31	3	33	28	11	94 (87-96)	55
B-Eos&FeNO-low	30	64	14	55	20	46	9	7	93 (88-102)	20
AwCOPD	2	24	20	93	61	6	23	30	97 (57-97)	30
Non-classified†	14	75	**8**	59	14	36	15	25	90 (80-103)	23
Multiple phenotypes:	43	72	23	64	14	**49**	**27**	**34**	**83 (74-95)**	34
AwObesity + others	48	71	23	65	16	48	**25**	**30**	**81 (74-92)**	26
B-Eos-high + others	19	75	20	65	14	**51**	**31**	**39**	**80 (69-91)**	**42**
FeNO-high + others	10	68	16	66	17	51	**30**	**37**	**81 (74-92)**	38
B-Eos&FeNO-low + others	9	78	33	59	18	42	18	25	90 (74-100)	21
AwCOPD +	5	61	22	61	20	**66**	**38**	**50**	**74**	**63**

others									**(63-86)**	
Specific combinations of phenotypes										
B-Eos-high + AwObesity	8	77	25	62	**8**	46	27	**32**	85 (77-95)	32
AwObesity+ B-Eos&FeNO-low	7	79	30	64	13	35	16	18	94 (77-101)	9
B-Eos-high + FeNO-high	5	73	19	66	**6**	44	**35**	41	80 (69-90)	**51**
B-Eos-high + FeNO-high + AwObesity	3	80	16	65	40	**67**	22	**33**	**79 (75-86)**	30
FeNO-high + AwObesity	2	**26**	8	67	12	47	27	32	93 (74-96)	9
B-Eos-high + AwCOPD	1	71	14	73	25	69	4	22	110 (90-110)	43
AwObesity + AwCOPD	1	**20**	19	67	0	71	12	14	**85 (83-85)**	14
AwCOPD+ B-Eos&FeNO-low	1	76	44	**20**	12	**84**	44	**71**	**66 (63-77)**	**85**

Legend: AwObesity: Asthma with obesity; AwCOPD: Asthma with COPD; ED: Emergency-department; FEV_1: Forced expiratory volume in 1 second; FEV_1/FVC: Forced expiratory volume in 1 second and functional vital capacity ratio; LLN: Lower limit of normality. Values are presented as weighted percentages and significant associations ($p<0.05$) between multiple and single phenotypes are presented in bold. *Subjects having only one of the 5 asthma phenotypes: AwObesity, B-Eos-high, FeNO-high, B-Eos&FeNO-low or AwCOPD. † Subjects with non-single and non-multiple phenotypes. ‡Subjects having at least one of the other asthma phenotypes. § Presented as median (Q1-Q3).

Table 8. Weighted percentages and comparisons of asthma-related outcomes among subjects with a single asthma phenotype versus: non-classified, and specific combinations of asthma phenotypes

Weighted %	Total	Asthma attack	Asthma-related ED	≥2 asthma symptoms	Work/school absenteeism	Asthma medication		Lung function		
						≥1 reliever medication	≥2 controller medication	FEV_1 <LLN	FEV_1% predicted§	FEV_1/FVC <LLN
Single phenotypes *	49	68	22	59	15	40	18	18	92 (82-99)	28
AwObesity	31	61	27	59	15	29	15	18	92 (84-111)	16
Type 2-high†	43	77	23	59	11	46	25	25	90 (80-98)	41
B-Eos&FeNO-low	24	64	14	55	20	45	9	7	93 (88-102)	20
AwCOPD	2	24	20	93	61	11	24	20	97 (57-97)	30
Non-classified‡	13	75	**8**	59	14	36	15	25	90 (80-103)	23
Multiple phenotypes ||	38	71	24	65	19	**51**	**27**	**32**	**85 (74-96)**	30
AwObesity+ others	26	71	23	67	20	49	26	**30**	**85 (75-96)**	26
Type 2-high+ others	19	70	20	68	21	**54**	**32**	**37**	**82 (74-93)**	36
B-Eos&FeNO-low+ others	9	78	33	59	18	42	18	25	90 (74-100)	21
AwCOPD+ others	6	64	23	66	18	**66**	**43**	**50**	**74 (63-86)**	**63**
Specific combinations of phenotypes										
Type 2-high+	15	70	22	66	22	52	25	**32**	**83**	28

AwObesity									**(76-94)**	
Type 2-high+ AwCOPD	2	74	24	**86**	11	**69**	**47**	18	90 (81-110)	**52**

Legend: AwObesity: Asthma with obesity; AwCOPD: Asthma with COPD; ED: Emergency-department; FEV_1: Forced expiratory volume in 1 second; FEV_1/FVC: Forced expiratory volume in 1 second and functional vital capacity ratio; LLN: Lower limit of normality. Values are presented as weighted percentages and significant associations ($p<0.05$) between multiple and single phenotypes are presented in bold. *Subjects with only one of the 4 asthma phenotypes: AwObesity, Type 2-high, B-Eos&FeNO-low or AwCOPD. †Type 2-high (B-Eos-high or FeNO-high), ‡ subjects with non-single and non-multiple phenotypes. ||Subjects having at least one of the other asthma phenotypes. § Presented as median (Q1-Q3)

Table 9. Multivariable logistic regression models between each asthma-related outcome and having multiple asthma phenotypes, adjusted for co-variables

	Asthma attack aOR (95% CI)	**Asthma-related ED** aOR (95% CI)	**≥2 asthma symptoms** aOR (95% CI)	**Work/school absenteeism** aOR (95% CI)	**≥1 reliever medication*** aOR (95% CI)	**FEV_1/FVC <LLN** aOR (95% CI)
Multiple vs. single phenotype	1.27 (0.78-2.06)	1.17 (0.74-1.86)	1.26 (0.76-2.11)	0.79 (0.37-1.68)	1.55 (0.97-2.49)	1.74 (0.94-3.24)
Female	1.34 (0.87-2.07)	**2.05 (1.08-3.90)**	**1.96 (1.16-3.31)**	1.35 (0.66-2.79)	0.88 (0.53-1.49)	**0.53 (0.31-0.90)**
Age ≥ 40 yrs	0.80 (0.50-1.27)	0.91 (0.59-1.41)	0.94 (0.55-1.60)	1.09 (0.60-2.00)	1.08 (0.67-1.73)	1.50 (0.90-2.49)
Caucasian vs. other	0.89 (0.54-1.46)	**0.38 (0.23-0.64)**	0.72 (0.46-1.15)	0.76 (0.39-1.50)	1.03 (0.65-1.63)	0.92 (1.16-3.51)
Current smoker vs. non-/ex-smokers	0.89 (0.55-1.43)	1.60 (0.94-2.72)	1.22 (0.69-2.14)	0.82 (0.37-1.83)	**1.95 (1.35-2.83)**	**2.02 (1.16-3.51)**
Rhinitis	0.96 (0.59-1.57)	0.77 (0.38-1.56)	0.88 (0.50-1.52)	0.70 (0.32-1.53)	0.66 (0.42-1.04)	1.04 (0.62-1.75)
Goodness-of-fit test χ^2 (p-value)	0.59 (0.80)	0.59 (0.81)	1.33(0.22)	1.27 (0.25)	1.14 (0.33)	3.05 (0.002)

Legend: CI: confidence interval; aOR: adjusted odds ratio; ED: emergency-department; FEV_1/FVC: Forced expiratory volume in one second and forced vital capacity ratio; LLN: Lower Limit of Normal; χ^2: Chi-square. All asthma-related outcomes were analyzed separately and treated as dependent variables. The aOR values with $p<0.05$ are presented in bold. *Short-acting β_2-agonist and/or anticholinergic.

A lower proportion of subjects reporting asthma attacks was observed, in subjects with AwObesity and either FeNO-high (26%) or AwCOPD (20%), compared to those with a single phenotype (67%) (Table 7). Subjects with concomitant AwCOPD and B-Eos&FeNO-low phenotypes had the lowest proportion of ≥ 2 asthma symptoms (20%) but had the highest proportion of using ≥ 1 reliever medication (84%) as well as having FEV_1< LLN (71%).

In multivariate regression analysis, adjusting for covariables, having multiple phenotypes was significantly associated with using ≥ 2 controller medications (aOR, 95% CI 2.03, 1.16–3.57), and having reduced FEV_1 (3.21, 1.73–5.94) (Table 10). However, no associations were seen with asthma attacks, asthma-related ED, ≥ 2 asthma symptoms, work/school absenteeism, use of reliever medication or FEV_1/FVC < LLN (Table 9).

Table 10. Regression models with significant associations between having multiple asthma phenotypes and asthma-related outcomes, adjusted for co-variates

	≥2 controller medications		**FEV_1<LLN**	
	aOR	95% CI	aOR	95% CI
Multiple vs. single phenotype	**2.03**	**1.16-3.57**	**3.21**	**1.74-5.94**
Female gender	1.39	0.77-2.50	1.51	0.81-2.81
Age ≥40yrs	**3.01**	**1.52-5.95**	**2.55**	**1.29-5.03**
Caucasian vs. others	1.38	0.86-2.23	1.37	0.78-2.42
Current smoker vs. non-/ex-smokers	1.02	0.52-2.02	**2.01**	**1.21-3.33**
Rhinitis	1.08	0.57-2.16	0.94	0.54-1.63
Goodness-of-fit test χ^2 (p-value)	0.86 (0.56)		0.80 (0.61)	

Legend: Multivariate logistic regression models adjusted for gender, age, race, current smoking and rhinitis. The aOR values with p<0.05 are presented in bold. FEV_1: Forced expiratory volume in 1 second; LLN: Lower limit of normal; CI, confidence interval; aOR, adjusted odds ratio; χ^2, Chi-square goodness-of-fit.

Furthermore, subjects aged ≥ 40 years-old, had significantly higher odds of using ≥ 2 controller medications and having FEV_1 < LLN predicted, compared to those < 40 years-old, adjusted for covariates (Table 10). Being a current smoker was significantly associated with using ≥ 1 reliever medication (1.95, 1.35–2.83) and with reduced lung function: FEV_1< LLN predicted (2.01, 1.21–3.33) and FEV_1/FVC < LLN (2.02, 1.16–3.51) and not associated with any other asthma-related outcomes (Table 9 and Table 10).

The association between having concomitant phenotypes and using multiple controller medications was consistent when considering oral corticosteroids (OCS) separated from other controller medications (1.87, 1.09–3.21) (data not shown).

We also analyzed the potential bias of controller medications in the phenotype classification, particularly in the B-Eos-high and FeNO-high phenotypes (Figure 7).

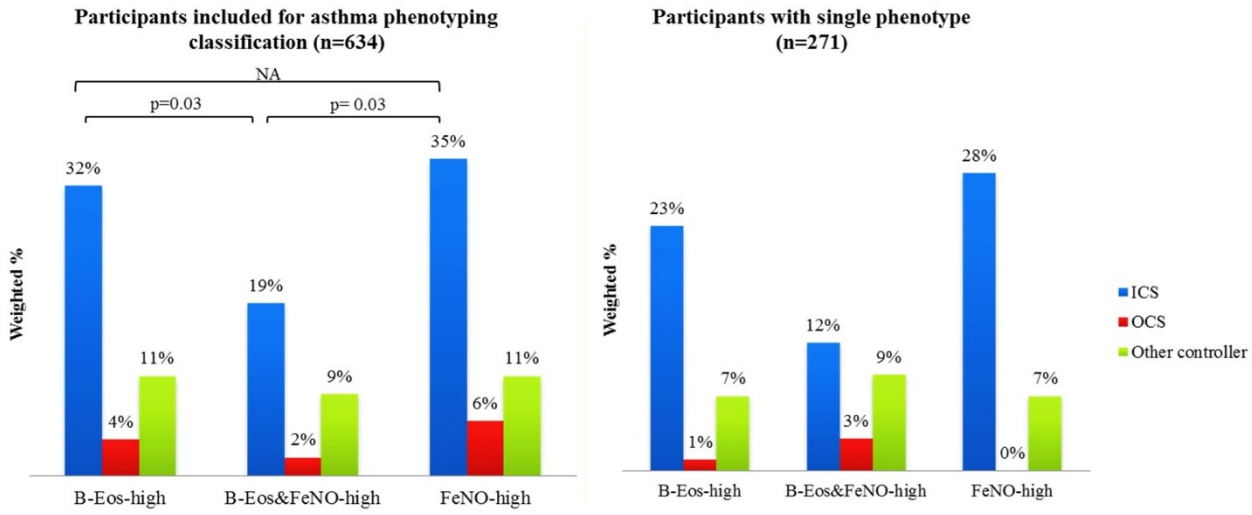


Figure 7. Proportions of subjects having asthma controller medications stratified into the different phenotypes. Left: all participants included for asthma phenotype classification. Right: and only in those with a single phenotype. NA: Non-applicable (as some participants had both B-Eos-high and FeNO-high asthma phenotypes)

No significant differences in asthma related treatment were found between the phenotypes, with exception for a higher proportion of patients treated with ICS within the FeNO-high and B-Eos-high phenotypes compared to those with B-Eos&FeNO-low phenotype (p = 0.03). When restricting to subjects with a single asthma phenotype no significant differences were found.

Moreover, sensitivity analyses showed that the proportion of total overlap (weighted 53%), and the associations between having multiple phenotypes and asthma outcomes were similar when imputing all missing values (data not shown). The goodness-of-fit test revealed adequate fitting for all regression models, except when using $FEV_1/FVC<LLN$ as dependent variable (Table 9) and no statistically significant interactions between co-variables were observed.

3.2 Study II

Amaral, R., Pereira, A. M., Jacinto, T., Malinovschi, A., Janson, C., Alving, K., & Fonseca, J.A.

"Comparison of hypothesis- and data-driven asthma phenotypes in NHANES 2007-2012: the importance of comprehensive data availability".

Clinical and Translational Allergy. 2019; 9:17.

The aim of study II (Appendix III) was to compare previously defined hypothesis-driven asthma phenotypes with data-driven asthma phenotypes derived by applying latent class analysis, in a sample of adults representative of the US general population.

Methods

Study setting and participants

We have included subjects that participated in the NHANES study, a nationally representative survey of the civilian, non-institutionalized US population performed with the aim of gathering data regarding health and nutritional status. Protocols were approved by the National Center for Health Statistics Research Ethics Review Board and all participants gave written informed consent.

Data from three NHANES surveys was used (n=30,442). We included adults (≥18 years old) with current asthma (n=1,059), defined by a positive answer to the questions (Amaral et al., 2018a): “Has a doctor ever told you that you have asthma?” together with “Do you still have asthma?”, and either “wheezing/whistling in the chest in the past 12 months” or “asthma attack in the past 12 months.”

Variables

Anthropometric and demographic characteristics, such as age, gender, BMI, and smoking status were analysed, as well as B-Eos count, FeNO and spirometric parameters. FeNO and spirometry were performed following ATS/ERS recommendations (Miller, 2005; Silkoff, 2005). Basal predicted values of FEV_1 and FVC were calculated (Hankinson et al., 1999; Hankinson et al., 2010) and abnormal values were defined as being below the LLN (Stanojevic et al., 2008).

The main questions used to define the asthma-related questions were collected from the structured clinical interview assessment. Variables were defined by the following questions:

- *Smoking status*: current smoker defined as smoking at least one cigarette every day for one year; ex-smokers reported having quit smoking for more than one month; non-smokers reported neither smoking nor ex-smoking.
- *Self-reported chronic bronchitis/emphysema*: Positive answer to either “Has a doctor ever told you/SP that you had chronic bronchitis?” or “Has a doctor ever told you/SP that you had emphysema?”

- *Age of chronic bronchitis/emphysema diagnosis*: defined with the answer to the question “How old were you when you were first told you had chronic bronchitis/emphysema?”
- *Age of asthma onset*: defined with the answer to the question “How old were you when you were first told you had asthma?”
- *Wheezing attack*: defined if at least one was reported to the question “In the past 12 months, how many attacks of wheezing or whistling have you?”
- *Wheezing with exercise*: Positive answer to “In the past 12 months, has your chest sounded wheezy during or after exercise or physical activity?”
- *Sleep disturbance by wheezing*: defined if at least one was reported to the question “In the past 12 months, how often, on average, has your sleep been disturbed because of wheezing?”
- *Limit activity by wheezing*: defined if at least a little amount was reported to the question to “During the past 12 months, how much did you/SP limit your usual activities due to wheezing or whistling?”
- *Absenteeism by wheezing*: defined if at least 1 to 7 days option was reported to the question “During the past 12 months, how many days of work or school did you miss due to wheezing or whistling?”
- *Asthma-related emergency department visits*: Positive answer to “During the past 12 months, have you had to visit an emergency room or urgent care center because of asthma?”
- *Hay fever*: Positive answer to “During the past 12 months, have you had an episode of hay fever?”
- *Oral corticosteroids use*: Positively answer to “In the past month, have you used or taken medication for which a prescription is needed?” and then reported OCS use.

Hypothesis-driven asthma phenotypes

The analysis based on the report of smoking status, presence of obesity and inflammatory markers enabled the definition of five asthma phenotypes (Amaral et al., 2018a): B-Eos-high asthma phenotype, if B-Eos ≥ 300/mm^3; FeNO-high asthma, if FeNO ≥ 35ppb; B-Eos&FeNO-low asthma, if B-Eos < 150/mm^3 and FeNO < 20ppb; AwObesity, if BMI ≥ 30 kg/m^2; and AwCOPD, if subjects had self-reported chronic bronchitis/emphysema with age of diagnosis ≥40 years and being either a current or an

ex-smoker (ever smoked). Subjects were considered as “non-classified” if they did not meet the criteria for any of the defined asthma phenotypes. Additionally, to account for individuals with probable co-existence of asthma and COPD and minimize age as a confounding variable, we conducted the analysis considering two age groups: <40 and ≥40 years old (Amaral et al., 2018a).

Data-driven asthma phenotypes

LCA was used to identify asthma phenotypes in an unsupervised manner, i.e. without the need for historical or a priori assumptions (data-driven approach). Each participant was assigned to one class for each measure with the highest posterior class membership probability. The most appropriate number of classes was determined by examining commonly used criteria, including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the sample size-adjusted BIC, the Lo-Mendell-Rubin Likelihood Ratio Test, and the entropy (Muthén & Muthén, 2012).

Two models for “current asthma” were developed (Table 11): Model 1 was based on the 4 variables previously used to define the hypothesis-driven asthma phenotypes (BMI ≥30kg/m^2, ever-smoking status, FeNO ≥35ppb, B-Eos ≥300/mm^3) (Amaral et al., 2018a); and in Model 2, we have added to the former 4 variables, sex, early asthma onset (<16 years old), wheezing-related questions (presence/absence of at least one wheezing attack, wheezing with exercise, sleep disturbance by wheezing, limit activity by wheezing, absenteeism by wheezing), asthma-related ED visit in the previous 12 months, FEV_1/FVC <LLN, FEV_1<LLN, and self-reported hay fever.

Additionally, to explore the results in different “asthma populations”, we’ve developed two other models using similar variables. For the “ever asthma” subgroup (model 3) we included subjects with a positive answer to “Has a doctor ever told you that you have asthma?” (n=2,611); and for the “difficult asthma” (model 4) we included subjects with poor asthma-related outcomes, defined as current asthma plus, at least, one of the following: asthma-related ED visit, FEV_1<LLN, or oral corticosteroids use in the past 30 days (n=673) (Table 11).

Latent class models were derived independently for each age group, using the same variables, and a secondary analysis without stratifying by age was done on the three asthma subgroups.

Table 11. Description of the LCA-models and the respective included variables in the different "asthma populations". In gray is the selected model used to compare with the hypothesis-driven asthma phenotypes

	Variables included in the model	Number of LCA classes (p-value*)		
		<40 yrs old	≥40 yrs old	Without age stratification
Model 1 "Current asthma" † n= 1,059	- BMI ≥30kg/m^2 (Y/N) - Ever smoked (Y/N) - FeNO ≥35ppb (Y/N) - B-Eos ≥300/mm^3 (Y/N)	1 (N/A)	1 (N/A)	1 (N/A)
Model 2 "Current asthma" † n= 1,059	- BMI ≥30kg/m^2 (Y/N) - Ever smoked (Y/N) - FeNO ≥35ppb (Y/N) - B-Eos ≥300/mm^3 (Y/N) - Sex (M/F) - Early asthma onset (Y/N) - Wheezing-related questions (Y/N): at least one wheezing attack, wheezing with exercise, sleep disturbance/limit activity/absenteeism by wheezing - FEV_1/FVC <LLN (Y/N) - FEV_1<LLN (Y/N) - Asthma-related ED visit (Y/N) - Hay fever (Y/N)	2 (0.003)	2 (0.04)	2 (<0.001)
Model 3 "Ever asthma" ‡ n= 2,611	- BMI ≥30kg/m^2 (Y/N) - Ever smoked (Y/N) - FeNO ≥35ppb (Y/N) - B-Eos ≥300/mm^3 (Y/N) - Sex (M/F) - Early asthma onset (Y/N) - Wheezing-related questions (Y/N): at least one wheezing attack, wheezing with exercise, sleep disturbance/limit activity/absenteeism by wheezing	2 (<0.001)	2 (0.001)	2 (<0.001)

	- FEV_1/FVC <LLN (Y/N) - FEV_1<LLN (Y/N) - Asthma-related ED visit (Y/N) - Hay fever (Y/N)			
Model 4 "Difficult asthma" ¥ n= 673	- BMI ≥30kg/m^2 (Y/N) - Ever-smoking status (Y/N) - FeNO ≥35ppb (Y/N) - B-Eos ≥300/mm^3 (Y/N) - Sex (M/F) - Early asthma onset (Y/N) - Wheezing-related questions (Y/N): at least one wheezing attack, wheezing with exercise, sleep disturbance/limit activity/absenteeism by wheezing; - Hay fever (Y/N)	1 (N/A)	1 (N/A)	1 (N/A)

Legend: N/A: not applicable; Y/N: Yes/No; BMI: body mass index; FeNO: Fractional exhaled nitric oxide; B-Eos: Blood eosinophils count; FEV1: Forced Expiratory Volume in the first second; FVC: Forced vital capacity; LLN: Lower limit of normality; ED: Emergency department. * p-value obtained by the Lo-Mendell-Rubin Likelihood Ratio Test. † Current asthma defined as a positive answer to the questions: "Has a doctor ever told you that you have asthma?" together with "Do you still have asthma?", and either "wheezing/whistling in the chest in the past 12 months" or "asthma attack in the past 12 months." ‡ Ever asthma defined as a positive answer to "Have you ever had asthma?" ¥ Difficult asthma defined as current asthma plus, at least, one of the following: asthma-related ED visit in the previous 12 months, FEV_1<LLN, or LLN, or oral corticosteroids use in the past 30 days.

Statistical analysis

All analyses considered the complex multistage sampling and 6-year sampling weights provided by the NHANES documentation (CDC & National Center for Health Statistics, 2013). LCA was performed with MPlus (version 6.12), that considered the complex survey design of NHANES when performing LCA-modelling. All other analysis was performed in Stata/IC 15.1 (Stata Corp, College Station, TX, USA). A p-value <0.05 was considered statistically significant.

Results

We included 1,059 adults with current asthma. The weighted proportions of the previously defined hypothesis-driven asthma phenotypes, according to age groups (<40 and ≥40 years old) were, respectively: 42% and 53% with AwObesity; 34% and 37% with B-Eos-high asthma; 26% and 21% for B-Eos&FeNO-low; 18% and 19% with FeNO-high asthma; and 19% AwCOPD, in the older group (Amaral et al., 2018a). In addition, 17% and 12% of the individuals in the <40 and ≥40 years old groups, respectively, were categorized as “non-classified”.

In Model 1, LCA was not able to differentiate any asthma subgroup among subjects with current asthma (Table 11). On the other hand, by adding more asthma-related variables (Model 2), LCA identified a two-class model as the best solution for both age groups (Table 11 and Table 12). Classes A<40 years (n=290;75%) and A≥40 years (n=494;73%) had marked predominance of highly symptomatic asthma subjects, with poorer lung function, compared to classes B<40 years (n=96;25%) and B≥40 years (n=179;27%), respectively (Table 12). Regarding inflammatory markers, the proportion of patients with high levels of B-Eos and FeNO was not significantly different between classes, both in the younger group (p= 0.99 and p=0.82, respectively) and in the older group (p=0.57 and p=0.53).

Figure 8 shows that the distribution of the hypothesis-driven phenotypes is similar (p>0.05) in both classes identified by LCA regardless age group.

Additionally, LCA identified 2 classes on the models for “ever-asthma” and “current asthma” without stratifying by age, but not for the difficult-asthma sub analysis where no subgroup was identified (Table 11).

Table 12. Proportions of each variable according to the LCA-classes identified in Model 2 (subjects with current asthma, n=1,059). Variables are ordered by the highest mean difference between the 2 classes of each age group and each coloured box represents the prevalence of the variables within the class, ranging from 0% (light yellow) to 100% (red).

Class A<40 years n=290 (75%)	**Class B<40 years** n=96 (25%)	**Variables**	**Class A≥40 years** n=494 (73%)	**Class B≥40 years** n=179 (27%)
99%	0%	Limit activity by wheezing	65%	2%
76%	0%	Wheezing with exercise	68%	8%
64%	0%	Sleep disturbance by wheezing	63%	0%
66%	34%	Wheezing attack	99%	45%
29%	10%	FEV_1/FVC <LLN	34%	15%
19%	0%	Absenteeism by wheezing	16%	0%
28%	14%	ED by asthma	26%	11%
41%	59%	Female	66%	55%
20%	11%	FEV_1 <LLN	35%	19%
16%	25%	FeNO ≥35 ppb	21%	15%
34%	35%	B-Eos ≥300/mm^3	40%	30%
47%	41%	Ever smoked	61%	56%
26%	20%	Hay fever	44%	46%
43%	41%	BMI ≥30kg/m^2	55%	52%
51%	50%	Early asthma onset	25%	22%

Legend: FEV_1: Forced Expiratory Volume in the first second; FVC: Forced vital capacity; LLN: Lower limit of normality; ED: Emergency department; FeNO: Fractional exhaled nitric oxide; B-Eos: Blood eosinophils count; BMI: body mass index.

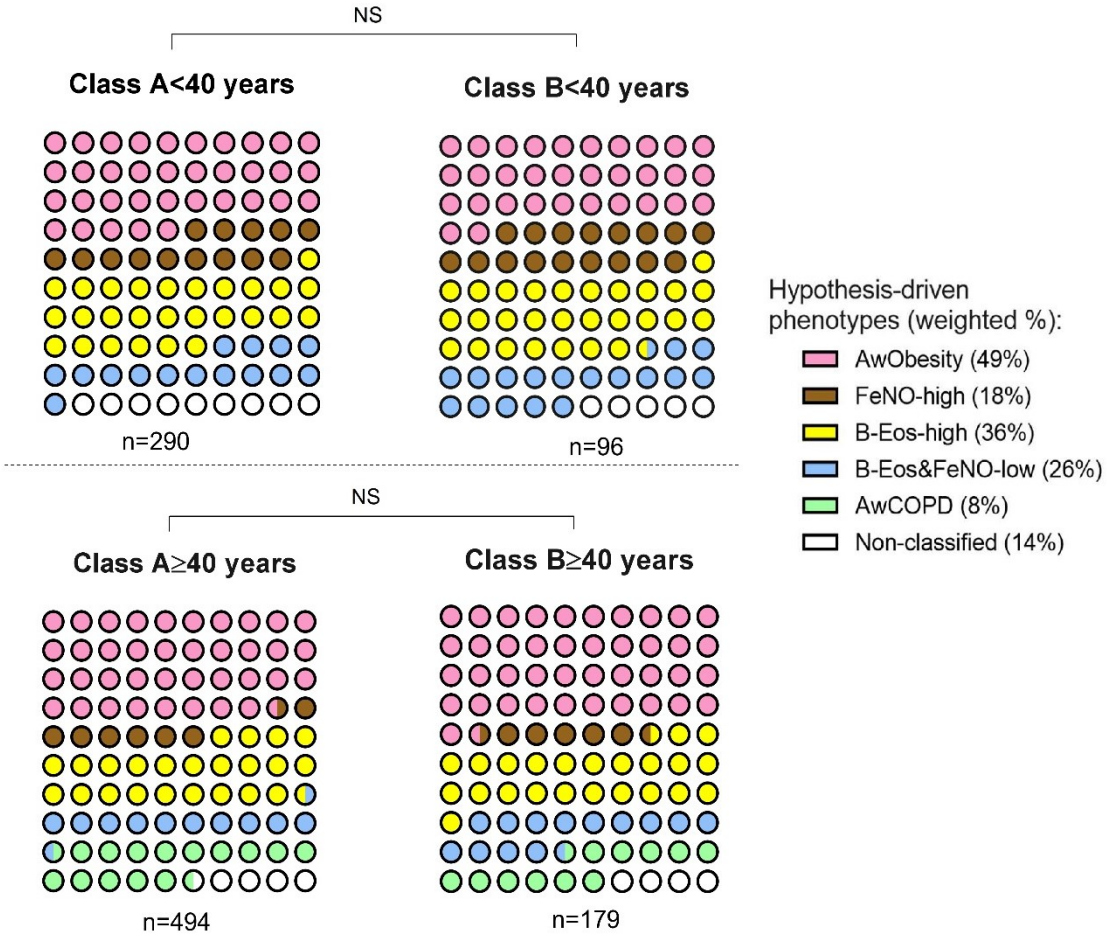


Figure 8. Distribution of the hypothesis-driven asthma phenotypes according to the data-driven classes identified in Model 2 (subjects with current asthma). Both Class A<40 and Class A≥40 are the phenotypes with more asthma-related symptoms and low lung function. No significant differences (p>0.05) were observed between the proportions of the hypothesis-driven within the data-driven phenotypes. NS: Non-significant

3.3 Study III

Amaral, R., Bousquet, J., Pereira, A.M., Araújo, L., Sá-Sousa, A., Jacinto, T., Almeida, R., Delgado, L. & Fonseca, J.A.

"Disentangling the heterogeneity of allergic respiratory diseases by latent class analysis reveals novel phenotypes".

Allergy. 2019; 74(4), 698-708.

The aim of study III (Appendix IV) was to identify distinct classes of allergic respiratory diseases using latent class analysis, in Portuguese adults from a general population sample and to explore the most relevant clinical variables that could be used to distinguish each class, using classification and regression tree analysis.

Methods

Sample and study design

This is a secondary analysis of a national and cross-sectional study conducted in the Portuguese general population, ICAR (Control and Burden of Asthma and Rhinitis) study (Sá-Sousa et al., 2019). All subjects that have been included in the first Portuguese National Asthma Survey (INAsma) (Sá-Sousa et al., 2012; Sá-Sousa et al., 2015) and that have expressed their willingness to participate in a clinical assessment, were eligible for ICAR study, along with their family members. Briefly, INAsma study was a nationwide, telephone interview study, aiming to evaluate asthma prevalence and control among the Portuguese population (Sá-Sousa et al., 2012; Sá-Sousa et al., 2015). It was based on a random sample of households within each municipality, using a list of landline phone numbers.

Participants were then screened by telephonic interview into one of the groups: self-reported diagnosis of asthma or/and rhinitis, and participants with no history of respiratory symptoms or diseases (Table 13), only to ensure that all the requirements of the sample size were fulfilled and not for disease diagnostic purposes. Additionally, an advertisement in the local media invited volunteers to participate in the clinical assessment regardless of previous diagnosis, to increase sample size (Figure 9).

Table 13. Distribution of the pre-screening categories according to the obtained LCA classes, in the ICAR participants

Self-reported diagnosis:	**Asthma and rhinitis** n=104	**Rhinitis only** n=268	**Asthma only** n=48	**Without asthma/rhinitis** n= 308
Class 1	3 (3)	21 (8)	3 (6)	155 (50)
Class 2	6 (6)	67 (25)	4 (8)	80 (26)
Class 3	15 (14)	23 (9)	12 (25)	27 (9)
Class 4	14 (14)	74 (28)	4 (8)	10 (3)
Class 5	18 (17)	53 (20)	11 (23)	32 (10)
Class 6	48 (46)	30 (11)	14 (29)	4 (1)

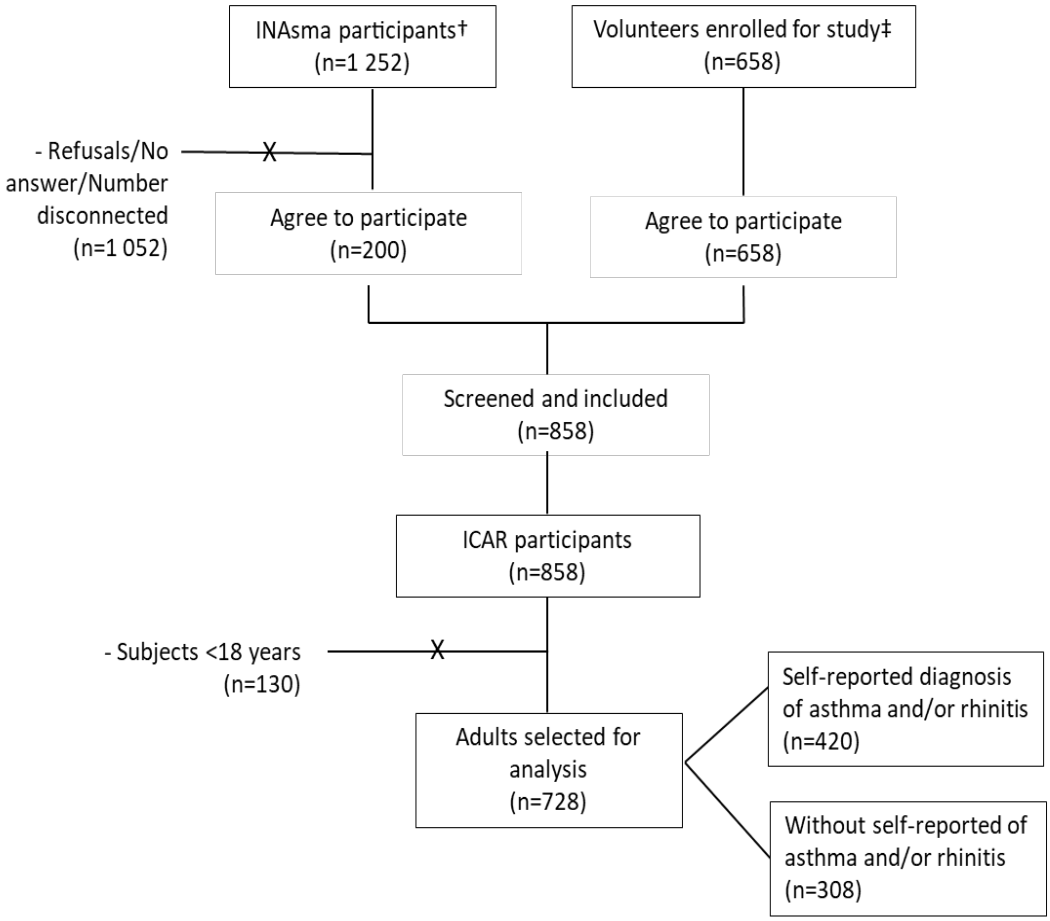


Figure 9. Diagram summarizing the inclusion of individuals. † Participants from the INAsma study that stated willingness to participate in a clinical assessment; ‡ Volunteers were family members of the INAsma patients or recruited by advertisement in the local media.

Data collection took place between 30th October 2012 and 12th July 2014, at the participants' communities, encompassed all districts of the continent, using a mobile diagnostic unit or at two allergy clinics in Lisbon and Porto, based on geographical convenience. It comprised anthropometric measurements, lung function, inflammation and allergy tests, a structured clinical assessment and standardized questionnaires. The data collection was part of a funded FCT project (PTDC/SAU-SAP/ 119192/2010).

ICAR study enrolled 858 participants, from 3 to 89 years-old, with and without self-reported asthma and/or rhinitis (Figure 9) and, for this analysis, we considered all adults (≥ 18 years old, n=728) (Table 13).

Ethical approval was obtained from the Hospital Ethics Committee (Comissão de Ética do Hospital São João, E.P.E) and national data protection committee (n.12372/2011). All participants gave written informed consent. The paper follows the STROBE guidelines for reporting observational studies (von Elm et al., 2008).

Variables

Demographic characteristics, such as age, gender, BMI and smoking status were analyzed. Data regarding nasal, bronchial and ocular symptoms, bronchial exacerbation and unscheduled medical visit, in the last 12 months, were collected by clinical interview, performed by a trained physician.

Variables were defined by the following questions:

- *Smoking status*: Non-smokers were considered if participants smoked <100 cigarettes in life-time; current smoker if responded “Every day”/ “Some days” to the question: “Do you now smoke cigarettes?”; and a response of “Not at all” was coded as former smoker.
- *Age at onset of the bronchial symptoms*: “How old were you when you begin to have asthma symptoms?”
- *Bronchial exacerbation* was defined as bronchial symptoms worsening, in the past 12 months, with/without the need for oral corticosteroid use, assessed in the clinical interview.
- *Unscheduled medical visit* was defined as having at least one emergency department and/or any unscheduled medical attendance due to bronchial exacerbation, in the previous 12 months.
- *Sneezing*: “During the past 12 months, have you had a problem with sneezing when you did not have a cold or the flu?”
- *Rhinorrhea*: “During the past 12 months, have you had a problem with runny nose when you did not have a cold or the flu?”
- *Nasal pruritus*: “During the past 12 months, have you had a problem with itchy nose when you did not have a cold or the flu?”
- *Nasal congestion*: “During the past 12 months, have you had a problem with blocked nose when you did not have a cold or the flu?”
- *Impaired sleep due to nasal symptoms*: “During the past 12 months, have these nose problems disturbed your sleep?”
- *Impairment in daily activities by nasal symptoms*: “During the past 12 months, have these nose problems interfere in your daily activities?”
- *Impairment in work/school by nasal symptoms*: “During the past 12 months, have these nose problems interfere in your work or school?”
- *Dyspnea*: “In the past 12 months have you had shortness of breath/dyspnea?”
- *Dyspnea at night*: “In the past 12 months have you woke up during the night due to shortness of breath/dyspnea?”

- *Wheezing*: “In the past 12 months have you had wheezing or whistling in your chest?”
- *Chest tightness*: “In the past 12 months have you had chest tightness?”
- *Watery eyes*: “In the past 12 months have you had watery eyes?”
- *Itchy eyes*: “In the past 12 months have you had itchy eyes?”

Nasal severity score was adapted from the ARIA severity score (Bousquet et al., 2018) and it was calculated using questions regarding impact of nasal symptoms on daily activities, work and sleep, ascribed with the score 1 if “Yes” and 0 if “No”. The nasal severity score was then categorized as “no/mild impairment” (ranging 0-2) and “severe impairment” (score=3).

Additionally, a diagnosis of rhinitis, asthma and other allergic diseases was established by an allergy specialist according to the structured and complementary exams. The use of asthma/allergy medication in the last 12 months, was also analyzed.

Measurements

The structured interview was conducted by a trained physician and included: physical examination; use of health resources and medications due to asthma/rhinitis; detailed personal and family medical history. The self-administered questionnaires have been previously reported (Sá-Sousa et al., 2019) and included detailed questions on recent or past respiratory symptoms and diseases control.

Allergic sensitization was assessed by skin-prick tests (SPT), following the GA2LEN recommendations (Bousquet et al., 2012). The standardized allergen panel included 28 allergens (Stallergenes Greer®, France) (Heinzerling et al., 2005), categorized into 6 groups: mites (*Dermatophagoides pteronyssinus*, *Lepidoglyphus destructor*, *Dermatophagoides farinae* and *Glycyphagus domesticus*), dog and cat epitheliums, tree (olive, plane, birch, cypress, pine, poplar, cork, oak, chestnut, hazel, alder), weed (*Parietaria judaica*, *Artemisia vulgaris*, *Plantago lanceolata*, *Ambrosia tryphida*, *Chaenopodium album* and *Urtica dioica*), grass pollens mixtures and molds (*Penicillum*, *Cladosporium*, *Aternaria alternata* and *Aspergillus fumigatus*). All SPTs included a positive (histamine 10 mg/ml) and negative control (saline/glycerol) (Heinzerling et al., 2013). Allergic sensitization was defined by a positive SPT with a mean wheal diameter ≥3 mm, for at least one of the 6 groups of allergens. Monosensitization and polysensitization were defined, respectively, as sensitization to only one and to two or more groups of allergens. Furthermore, sensitization data of 25 participants with current antihistamine medication were considered as missing.

Total IgE, serum ECP and B-Eos were obtained from blood sampling. Phadiatop® (Thermo Fisher Scientific, Uppsala, Sweden) was used as a screening test; if ≥0.35kU/l the sample was considered as Phadiatop-positive, and additional determinations were performed to assess individual allergen-specific IgE antibody concentrations.

Predicted values of basal FEV_1 and bronchodilator response were obtained by spirometry procedure, following the ATS/ERS recommendations (Hankinson et al., 1999; Miller, 2005). FEV_1 were considered abnormal if less than the LLN (Quanjer, Cole, Hall, & Culver, 2013). FeNO measurements were performed using NIOX MINO® (Aerocrine, Solna, Sweden), following the ATS/ERS criteria (Silkoff, 2005).

Biases

To reduce the risk of bias, several quality assurance measures were followed: research assistants performing the evaluations were blinded to the subject classification at screening; data validity was periodically verified soon after being collected and custom statistic algorithms were used to detect extreme, illogical and missing value and amendments to the protocol were done if necessary.

Statistical analyses

Categorical variables are presented as absolute frequencies and proportions. Continuous variables were presented according to their distributions. The socio-demographic and clinical variables of the classes derived from the LCA were described and compared using chi-square test, one-way analysis of variance and Kruskal-Wallis test with Bonferroni correction.

Mplus 6.12 (Los Angeles, CA: Muthén & Muthén) was used to conduct LCA analysis and R 3.3.3 (https://www.r-project.org/) to establish the classification model and build the respective decision tree, using the “rpart” and “DMwR” packages, respectively. All other analyses were performed using SPSS Statistics 25.0 (Armonk, NY: IBM Corp) and p-values <0.05 were defined as statistically significant.

Unsupervised analysis

LCA was applied to identify underlying unobserved (latent), mutually exclusive subgroups (classes) based on categorical manifest variables without the need for historical or *a priori* assumptions (Wang & Wang, 2012). Nineteen dichotomic variables (defined as Yes/No) were chosen regarding nasal, ocular and bronchial symptoms and the 6 groups of AS. The optimal number of classes resulting from the nineteen variables was determined by evaluating k classes versus k-1 classes sequentially, until

adding an additional class was no longer significantly improved, measured by Lo-Mendell-Rubin-Adjusted Likelihood Ratio test. The best model was determined by the largest entropy and the lowest BIC values.

Supervised analysis

Classification and regression tree (CART) analysis was performed to obtain the classification tree algorithm, using Gini impurity index and the Cost-Complexity pruning algorithm (Witten & Frank, 2005). This algorithm consists in generating a sequence of sub-trees of an overly large tree, then obtain individual reliable error estimates by cross-validation estimation procedure. The chosen result is within one standard error of the tree with best error estimate, to avoid overfitting and validating the classification tree internally. Variables that did not contribute significantly were automatically removed from the final model.

The algorithm allows to select the variables most likely to identify LCA-classes, which included parameters easily accessible in most outpatient settings. Variables included in the CART analysis were the following: demographic characteristics (gender, age, and BMI); presence of nasal (sneezing, rhinorrhea, nasal pruritus, nasal congestion), bronchial (dyspnea, dyspnea at night, wheezing, chest tightness) and ocular (watery eyes, itchy eyes) symptoms (Yes, if at least one symptom is present/No, without symptoms); nasal-related impairment: disturbed sleep (Yes/No), impairment of daily activities (Yes/No) and work/school (Yes/No), due to nasal symptoms; objective measures: number of groups allergic sensitization by SPTs, FEV_1 (categorized as <LLN and ≥LLN), and FeNO values.

Variable importance was given by Gini index (ranging 0-100%) (Sauve & Tuleau-Malot, 2014). Additionally, we randomly divided the dataset into a training (70%) and a test set (30%) to obtain a reliable estimate of the model's predictive performance. Cohen's kappa coefficient (kappa) was used to evaluate model performance for imbalanced datasets (Koch & Landis, 1977), using the "cohen.kappa" function of the "psych" package in R software.

Results

Sample characteristics

728 adults (63% female) were included, mean(sd) age of 43.9(15.2) years, 61% were non-smokers and 11% were taking ICS. Demographic and clinical characteristics of the 6 classes are shown in Table 14.

Latent class analysis

A six-class model was selected as the best solution for these data, with a significantly better fitting than a five-class model (p=0.013), and a non-significantly different fit from a seven-class model (p=0.363) (Table 15). Furthermore, entropy of the six-class model was 0.873, a good overall certainty in classification.

Figure 10 presents the probability of latent class membership for each of the six-class LCA model and Table 16 shows the stratification of the LCA-classes, according to clinical and allergic profiles. Average posterior probabilities were at least 89% for all classes, indicating a low chance of misclassification.

Two classes were characterized by non-allergic participants (>70% with negative SPT): classes 1 (n=182;25%) and 5 (n=114;16%). Class 1 had very low probability of having respiratory or ocular symptoms. Class 5 had a very high probability of having nasal, bronchial and ocular symptoms with nasal severe impairment (nasal severity score≥3).

Three classes were predominantly allergic (100% sensitization): classes 3 (n=77; 11%), 4 (n=96; 13%) and 6 (n=102; 14%). Class 3 had a high probability of nasal and ocular symptoms without severe nasal impairment (score ≤2). Classes 4 and 6 predominantly had high nasal and ocular symptoms with nasal impairment, differing by the absence (Class 4) or presence of bronchial symptoms (Class 6).

Class 2 (n=157; 22%) was characterized by participants with a very high probability of having nasal symptoms without severe nasal impairment, with a moderately increased probability of ocular symptoms, and 55% of them were allergic.

Table 14. Demographics and clinical characteristics of the 6 LCA-derived classes

	Total (n=728)	**Class 1** (n=182;25%)	**Class 2** (n=157;22%)	**Class 3** (n=77;11%)	**Class 4** (n=102;14%)	**Class 5** (n=114;16%)	**Class 6** (n=96;13%)	**p-value**
Female, n (%)	461 (63)	105 (58)	89 (57)	45 (58)	65 (64)	95 (83)	62 (65)	<0.001*
Age, mean (sd)	43.9 (15.2)	46.9 (16.5)	45.5 (16.4)	39.9 (14.8)	40.3 (12.9)	45.2 (13.4)	41.3 (13.2)	<0.001**
BMI, mean (sd)	25.9 (4.7)	26.6 (4.7)	25.1 (4.0)	25.8 (4.3)	25.2 (4.6)	25.8 (5.0)	27.1 (5.3)	0.003**
Age of bronchial symptoms onset, median (P_{25}-P_{75})	8.0 (3.0-20.0)	6.0 (1.0-13.0)	3.0 (2.0-18.0)	5.0 (3.0-7.5)	7.0 (2.0-12.0)	15.0 (3.0-25.0)	12.0 (4.0-30.0)	0.06***
Smoking status, n (%)								
Non-smoker	441 (61)	108 (59)	95 (60)	47 (61)	67 (66)	75 (66)	49 (51)	
Smoker	154 (21)	47 (26)	33 (21)	20 (26)	17 (17)	22 (19)	15 (16)	0.01*
Ex-smoker	133 (18)	27 (15)	29 (18)	10 (13)	18 (18)	17 (15)	32 (33)	
Packs-year, mean (sd)	5.9 (12.8)	8.3 (15.6)	6.5 (14.2)	5.0 (10.6)	3.6 (8.4)	4.2 (8.4)	5.7 (14.0)	0.03**
Current medication, n (%)								
ICS	81 (11)	3 (2)	6 (4)	14 (18)	5 (5)	18 (16)	35 (36)	<0.001*
Number of AS groups, median (P_{25}-P_{75})	2.0 (0-6.0)	0 (0-0.7)	1.0 (0-2.0)	4.0 (3.0-5.0)	4.0 (3.0-5.0)	0 (0-0)	4.0 (3.0-6.0)	<0.001***
Level of sensitization, n (%)								
Monosensitization†	100 (14)	37 (21)	37 (24)	0 (0)	0 (0)	22 (19)	4 (4)	<0.001*
Polysensitization‡	332 (46)	8 (4)	50 (32)	77 (100)	101 (100)	4 (3)	92 (96)	
Group of AS, n (%)								
Mites	336 (46)	31 (17)	55 (35)	73 (95)	77 (76)	19 (17)	81 (84)	<0.001*
Cat and dog epithelium	220 (30)	2 (1)	18 (11)	64 (83)	60 (59)	7 (6)	69 (72)	<0.001*
Molds	135 (19)	0 (0)	11 (7)	33 (43)	40 (40)	1 (1)	50 (52)	<0.001*
Pollens§	335 (46)	16 (9)	55 (35)	70 (91)	101 (100)	3 (3)	90 (94)	<0.001*
Lung function								
FEV_1% predicted, mean (sd)	97.4 (15.5)	99.4 (16.1)	99.6 (13.3)	95.3 (13.9)	98.9 (13.9)	97.2 (15.6)	90.1 (18.4)	<0.001**

FEV_1 <LLN, n (%)	69 (9)	16 (9)	8 (5)	6 (8)	9 (9)	8 (7)	22 (23)	<0.001*
Positive BD, n (%)	55 (8)	11 (6)	5 (3)	7 (9)	4 (4)	12 (11)	16 (17)	0.001*

Legend: BMI: body mass index; P25-P75: 25th Percentile-75th Percentile; ICS: Inhaled corticosteroids; AS: Allergen sensitizations; FEV_1: Forced Expiratory Volume in the first second; LLN: Lower limit of normal; BD: Bronchodilatation; † sensitization to 1 group of allergens; ‡ sensitization to 2 or more groups of allergens; § including tree, grass and weed sensitizations; * Chi-square test; ** One-way analysis of variance (ANOVA); *** Kruskal-Wallis test.

Table 15. Fit Statistics for LCA Models with 4 through 8 Classes

Number of classes:	4	5	6	7	8
AIC	11 815.7	11 646.1	11 498.9	11 456.1	11 363.5
BIC	12 141.6	12 054.6	11 990.1	12 029.9	12 020.0
Sample Adjusted BIC	11 916.1	11 772.0	11 650.3	11 633.0	11 565.9
Entropy	0.861	0.877	0.873	0.866	0.869
p-value LMR-LRT	0.001	0.014	0.013	0.363	0.305
Number of parameters estimated	71	89	107	125	143

Legend: AIC: Akaike Information Criteria; BIC: Bayesian Information Criteria; LMR-LRT: Lo-Mendell-Rubin-Likelihood Ratio Test

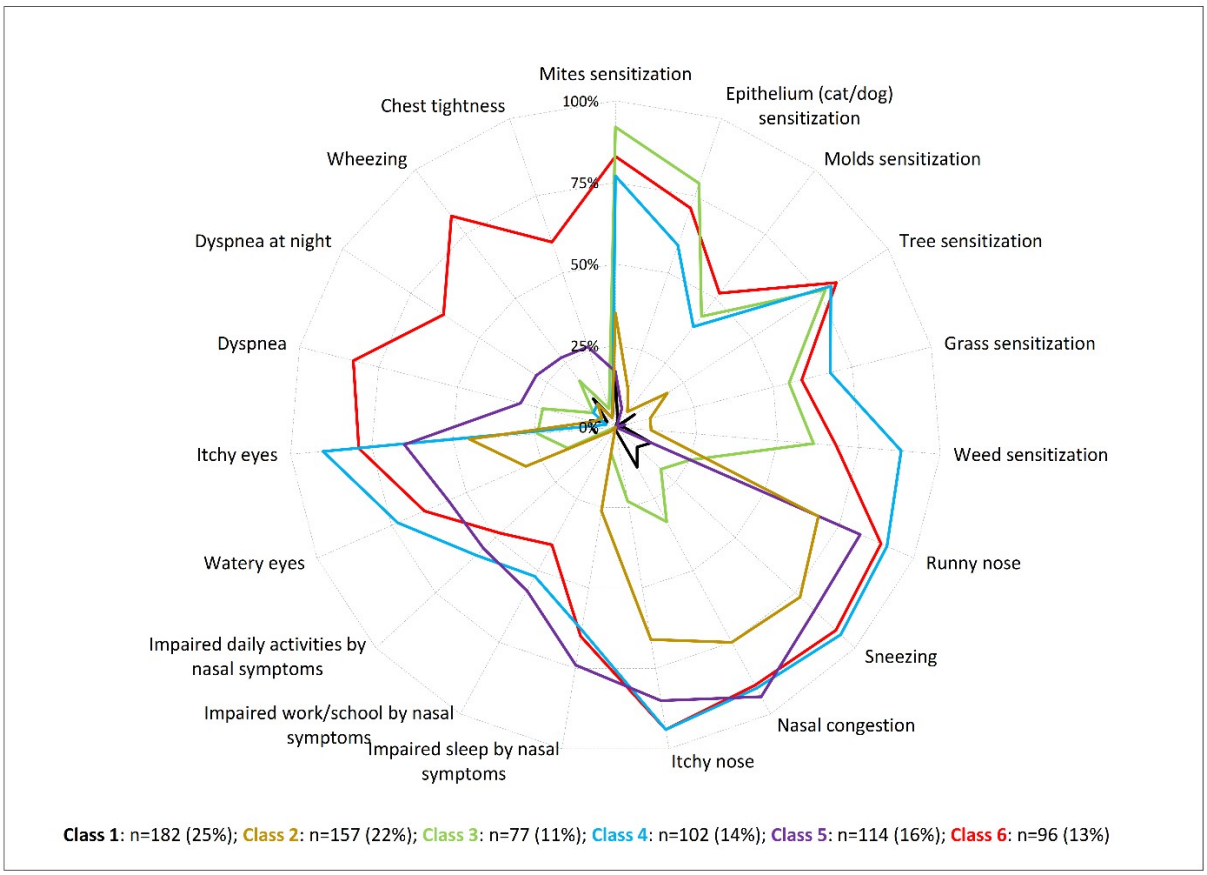


Figure 10. Proportions for the latent classes based on the estimated posterior probabilities

Latent class characteristics

There were significant differences among the 6 LCA-classes in all variables described in Table 14, except in age of bronchial symptoms onset (p=0.06). A female predominance across all classes was observed, particularly in Class 5 (83%).

Most participants in classes 3, 4 and 6 were polysensitized, with the more frequent allergic sensitization groups being: mites, pollens and cat/dog epithelia (Table 14 and Table 16). Moreover, half of the participants in Class 2 (n=87; 55%) were sensitized, particularly to mites (n=55; 35%) and classes 1 and 5 were mainly non-allergic.

Table 16. Classification/stratification of the LCA-classes, according to clinical and allergic profile

Class:	**1** n=182	**2** n=157	**3** n=77	**4** n=102	**5** n=114	**6** n=96
Rhinitis symptoms $	9	72	25	93	86	91
Asthma symptoms $	6	6	14	6	28	72
Ocular symptoms $	7	38	20	82	61	72
No. of nasal impairment*	0	0	0	3	3	1
Sensitization (%)						
No sensitization	75	45	0	0	77	0
Monosensitized	21	24	0	0	18	4
Polysensitized	4	32	100	100	3	96
Group of AS $						
Mite	17	35	92	77	17	83
Molds	0	6	43	39	1	52
Dog-cat epitheliums	2	12	79	59	6	71
Pollen £	4	14	64	78	1	69
Total IgE, geom mean	85.5	76.9	202.8	184.8	52.2	227.6
Specific IgE, median	0	0	3.6	7.4	0	13.2
Lung function						
FEV1, mean	99.4	99.6	95.3	98.9	97.2	90.1
Reversibility (%)	6	3	9	4	11	17
Current ICS use (%)	2	4	18	5	16	36
Smoking (%)	26	21	26	17	19	16
Healthcare use (%)						
Bronchial Exacerbations	2	2	15	7	16	54
Unscheduled medical visits	1	0	7	2	10	37

Legend: $: mean posterior probabilities (in %); * Any posterior probabilities >50 %; £: Pollens included weed, tree and grass. *Colour coding:* Green boxes indicated the more prevalent/frequent variables and yellow boxes the intermediate ones.

Regarding the presence of ocular symptoms (Figure 11), the proportion of participants with ocular symptoms is significantly higher in those with severe nasal impairment

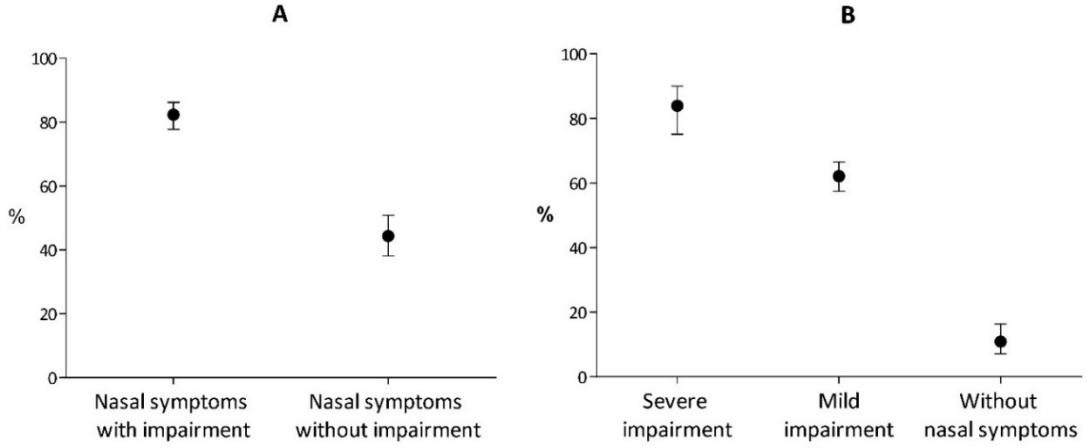


Figure 11. Proportions of participants with ocular symptoms, stratified by (A) the presence of nasal-related impairment and (B) ARIA severity score. Presence of nasal-related impairment defined as at least one of the following questions: impairment of daily activities by nasal symptoms, impairment of school/work by nasal symptoms and/or disturbed sleep due to nasal symptoms.

(score ≥3), compared to those without nasal impairment ($p<0.001$). Similarly, the rate of participants with ocular symptoms having severe nasal impairment was significantly higher when compared to those with mild nasal impairment (score ≤2) ($p<0.001$) or without nasal symptoms ($p<0.001$). Also, among participants without nasal symptoms, the proportion of ocular symptoms was very low (12%).

Classes 1 and 6 represented two extreme phenotypes: Class 1 was the mildest phenotype whereas Class 6 was the most severe, including participants with the lowest and the highest values of B-Eos, FeNO and total IgE and proportions of urgent medical care, respectively (Table 17 and Figure 12). Class 6 participants had significantly higher proportions of current use of ICS, abnormal lung function and positive bronchodilation as compared to other classes (Table 14). After Class 6, the non-allergic Class 5 had the second highest proportions of participants with current use of ICS (16%), positive bronchodilation (11%), and bronchial exacerbations, comparing to other classes.

An “Intermediate” phenotype was found. Class 3 had the lowest mean age, and when compared to other classes, had significantly higher proportions of participants with sensitization to indoor allergens, with 18% of them having current ICS medication (Table 14). Participants in Class 3 had a significantly higher proportion of bronchial exacerbations in the past 12 months, compared to classes 1 and 2 ($p<0.001$) (Figure 12).

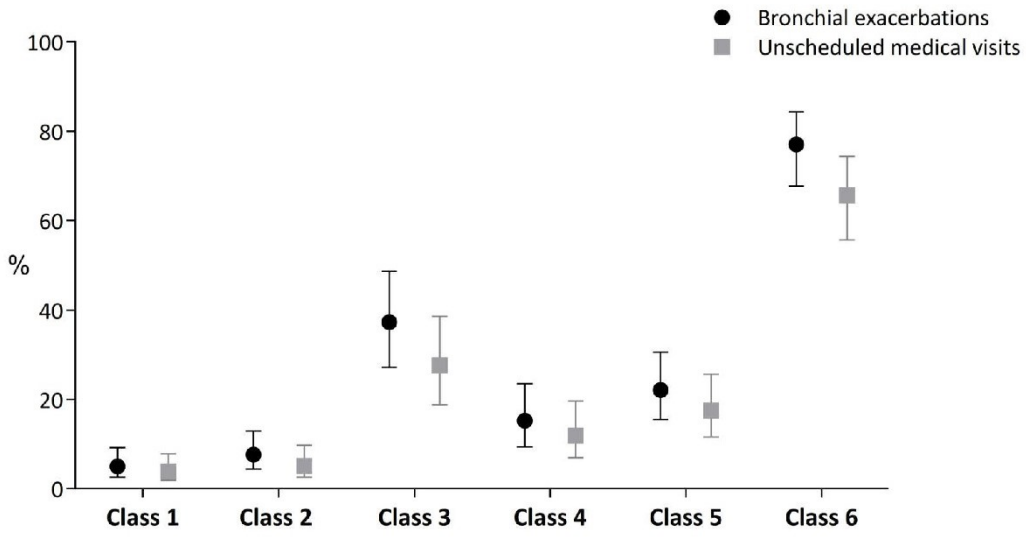


Figure 12. Proportions of participants with bronchial exacerbations and unscheduled medical visits, stratified by LCA-classes

Table 17. Distribution and comparison of FeNO and laboratory tests, according to each obtained LCA-classes

	Class 1 (n=182;25%)	**Class 2** (n=157;22%)	**Class 3** (n=77;11%)	**Class 4** (n=102;14%)	**Class 5** (n=114;16%)	**Class 6** (n=96;13%)	**Significant p-values†**
FeNO, geom mean (CI 95%)	17.9 (16.0-19.9)	18.1 (16.2-20.0)	24.9 (18.9-30.9)	26.5 (22.0-31.0)	19.3 (16.1-22.4)	35.1 (28.7-41.5)	C1/C4, C1/C6, C3/C4, C3/C6, C4/C5, C5/C6
S-ECP, geom mean (CI 95%)	13.2 (11.2-15.2)	14.5 (11.9-17.1)	13.2 (10.6-15.9)	15.1 (11.7-18.5)	12.0 (9.8-14.3)	16.8 (13.7-20.0)	N.S.
B-Eos, geom mean (CI 95%)	158.7 (133.6-183.8)	177.2 (152.3-202.0)	200.9 (169.0-232.7)	205.3 (174.7-236.0)	189.3 (161.8-216.8)	237.2 (197.4-277.1)	C1/C2, C1/C4, C1/C6
Phadiatop® sIgE, median (P_{25}-P_{75})	0 (0-0)	0 (0-1.6)	3.6 (0-16.5)	7.4 (1.9-23.2)	0 (0-0)	13.2 (0.6-31.3)	C1/C3, C1/C4, C1/C6, C2/C3, C2/C4, C2/C6, C3/C5, C4/C5, C5/C6
Total IgE, geom mean (CI 95%)	85.5 (52.4-118.7)	76.9 (54.7-99.2)	202.8 (138.5-267.0)	184.8 (111.0-258.6)	52.2 (36.2-68.2)	227.6 (170.0-285.3)	C1/C3, C1/C4, C1/C6, C2/C3, C2/C4, C2/C6, C3/C5, C4/C5, C5/C6

Legend: FeNO: Fractional exhaled nitric oxide; CI: Confidence interval; S-ECP: serum eosinophilic cationic protein; B-Eos: blood eosinophilic count; sIgE: specific IgE; C: Class; N.S.: non-significant. † significant if p-values<0.05, using Kruskal-Wallis test with Bonferroni correction.

Serum total IgE levels and Phadiatop® specific IgE were highest in sensitized groups (classes 3, 4 and 6) (Table 17). B-Eos values were significantly higher in classes 2, 4, and 6 comparing with Class 1. FeNO values were highest in classes 4 and 6, by comparison to classes 1, 3, and 5 ($p<0.03$) (Figure 13).

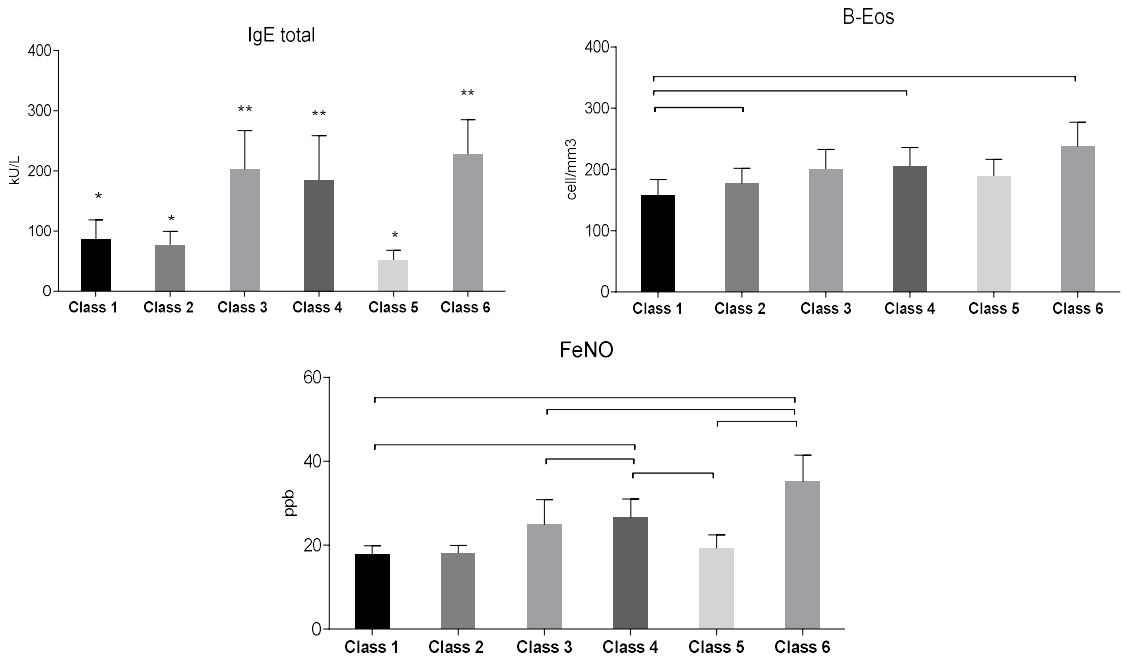


Figure 13. Comparison of total IgE, B-Eos and FeNO within each LCA-class. Different superscripts (* vs. **) and horizontal lines over bars indicate significant differences between classes ($p<0.05$), using Kruskall-Wallis test with Bonferroni correction. IgE: Immunoglobulin E; B-Eos: blood eosinophils count; FeNO: fractional exhaled nitric oxide.

Furthermore, medical diagnosis of rhinitis was common in classes 2, 4, 5 and 6 (Table 18), while asthma diagnosis was more frequent in Class 6 (75%), and less, but also prevalent, in classes 3 and 5 (35% and 31%, respectively). A high proportion of conjunctivitis was diagnosed in participants belonging to classes 4, 5 and 6. Class 1 participants had the lowest proportion of diseases medical-diagnosed, except for the proportion of other respiratory diseases (20%) (Table 18).

Table 18. Medical diagnosis of the 6 LCA-derived classes

	Class 1 n=182 (25%)	**Class 2** n=157 (22%)	**Class 3** n=77 (11%)	**Class 4** n=102 (14%)	**Class 5** n=114 (16%)	**Class 6** n=96 (13%)
Medical diagnosis, n (%)						
Asthma	10 (5)	13 (8)	27 (35)	12 (12)	35 (31)	72 (75)
Rhinitis	34 (19)	143 (91)	51 (66)	101 (99)	112 (98)	95 (99)
Conjunctivitis	19 (11)	80 (51)	22 (30)	93 (92)	84 (74)	79 (82)
Other resp. disease	36 (20)	12 (8)	10 (13)	4 (4)	6 (5)	4 (4)
Other allergic disease	2 (1)	0 (0)	2 (3)	0 (0)	0 (0)	0 (0)

CART analysis

Nine terminal nodes were formed in the classification tree (Figure 14), with a kappa (95% CI) =0.75 (0.72-0.79). Classification tree showed that ocular symptoms were the variable with the highest relative contribution to the model (37%), followed by number of AS groups (21%), having impairment of school/work by nasal symptoms (15%), presence of bronchial symptoms (13%), having impairment of daily activities (12%) and sleep disturbance due to nasal symptoms (3%).

On the right side of the tree (corresponding to participants sensitized to ≥3 AS groups), the presence of bronchial symptoms, distinguished Class 6 from all the others. When bronchial symptoms were absent, the presence/absence of impairment in daily activities and sleep disturbance by nasal symptoms, differentiated classes 3 and 4. Using a training (n=509) and a test set (n=219), the obtained CART algorithm was identical and similar kappa was also obtained (kappa [95% CI] = 0.73 [0.67-0.80]).

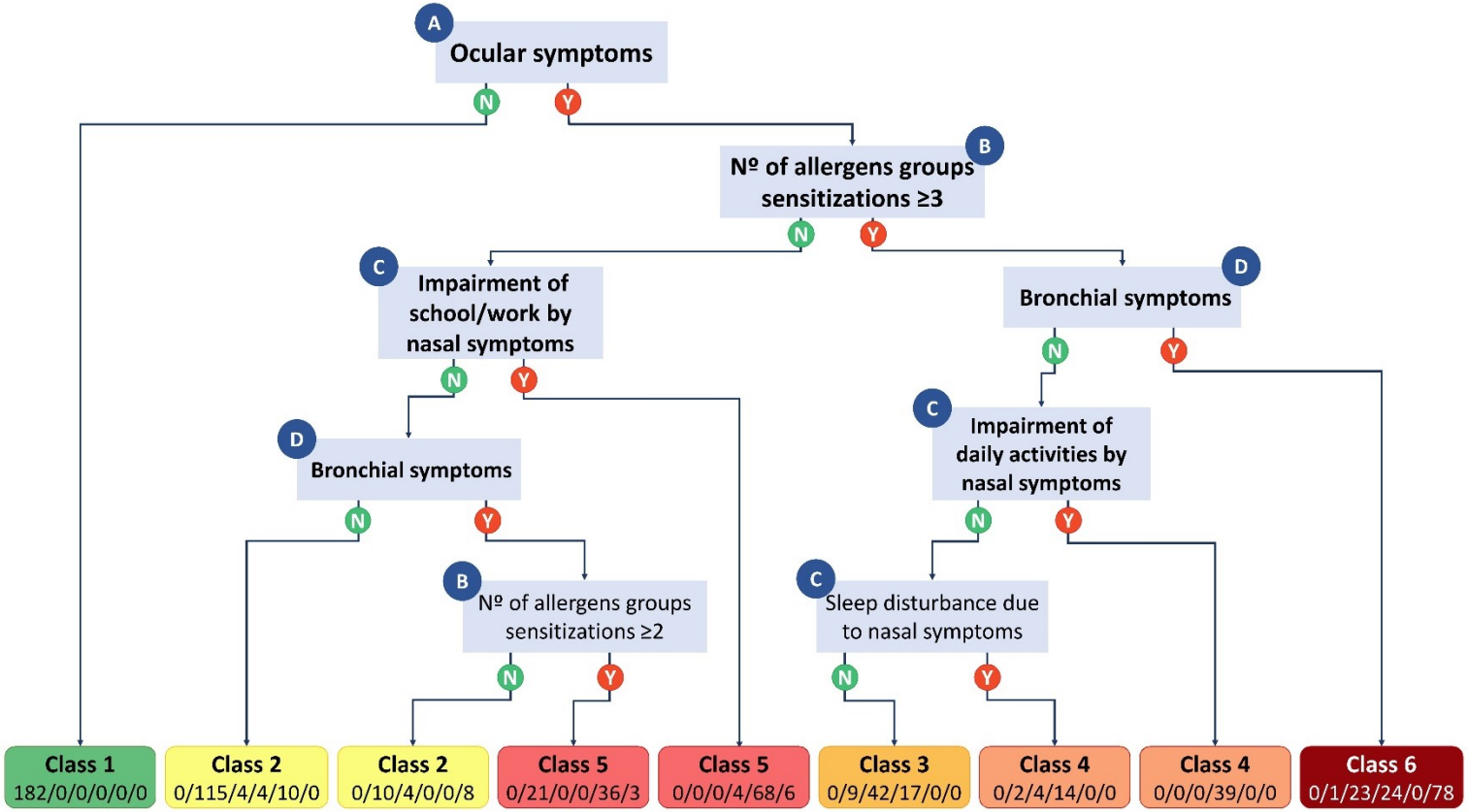


Figure 14. Classification tree algorithm generated by CART analysis using easily accessible parameters available in the clinical setting. (A) At least 1 ocular symptom: Watery eyes and/or Itchy eyes. (B) Mites, Cat/dog epitheliums, Tree, Grass, Weed, Molds. (C) Impairment due to nasal symptoms: Sneezing, Rhinorrhea, Nasal pruritus, Nasal congestion; (D) At least one bronchial symptom: Dyspnea, Dyspnea at night, Wheezing and/or Chest tightness. N: No; Y: Yes.

4. Discussion

This chapter provides a general discussion of the results from the three studies that compose this thesis. Its major findings are summarized and discussed, followed by its strengths, limitations and suggestions for future research. Finally, the concluding remarks are presented.

4.1 Major findings

A significant overlap of commonly reported, hypothesis-driven asthma phenotypes, was observed, in adults with current asthma in a large population sample, derived from the US NHANES surveys, with almost half of them having two or more concomitant phenotypes. Furthermore, having multiple asthma phenotypes, regardless of their classification, was associated with poorer asthma outcomes, namely the increased use of controller medication and reduced lung function.

In the same adults with current asthma, two data-driven phenotypes were identified using an unsupervised classification method, LCA, only differing on asthma-related symptoms and lung function.

The prevalent phenotypes overlap and the fact that only two data-driven phenotypes were identified, suggests that the clinical and physiological variables commonly used to subdivide asthma seem to be insufficient to differentiate specific asthma phenotypes in this population, regardless of using data-driven or hypothesis-driven approaches.

This suggestion was then corroborated when studying the comprehensive disease features available in the ICAR study, conducted in the Portuguese general population. Novel data-driven phenotypes of allergic respiratory diseases with co-occurrence of ocular, nasal and bronchial symptoms, and prone to exacerbations were revealed. Also, for the first time, ocular symptoms *per se* were used in LCA and were ranked, by a classification tree algorithm, as the most relevant variable to differentiate allergic respiratory diseases phenotypes.

In this work, we analysed multiple diseases/conditions not only using data-driven methods but also in individuals recruited directly from the general population rather than health services settings, unlike most studies that usually analyse severe disease phenotypes or data derived from tertiary care alone (Lefaudeux et al., 2017; Moore et al., 2010; Wu et al., 2014). Moreover, to our knowledge, this is the first time CART analysis is studied in a population-based study of ARD, using parameters easily accessible in most outpatient settings, with a potential to be further validated and applied in a routine clinical practice.

4.2 Interpretation of the results

In Study I five hypothesis-driven asthma phenotypes frequently used in the literature were defined (Bousquet et al., 2016), and a high overlap was shown between them. This finding was similar to the Tran et al. study (2016), that used datasets from previous NHANES surveys to evaluate the overlap of asthma phenotypes. However, the latter study focused only on allergic asthma phenotypes, based on IgE levels, and was therefore limited to the 2005–2006 survey that lacks data on FeNO. Other distinct subgroups of asthma phenotypes are increasingly being reported due to its characteristics of steroid therapy resistance and lack of inflammatory markers, e.g. subjects with asthma without evidence of type-2 inflammation (Th2-low phenotype) (Fahy, 2015; Robinson et al., 2017); obese asthmatic subjects (obesity-related asthma phenotype) (Gibeon et al., 2013); and patients with ACOS (Christenson et al., 2015). Therefore, in Study I, a broader analysis of phenotypes that included not only the eosinophil- and FeNO-based phenotypes (B-Eos-high, FeNO-high, and B-Eos&FeNO-low), but also other phenotypes not defined by biomarkers (AwObesity, AwCOPD) was found using a much larger dataset (six years NHANES survey).

Interestingly, there is a significant additive effect of having more than one asthma phenotype (regardless the specific combination) in abnormal FEV_1 and some asthma-related outcomes, such as using more controller medications, supporting the view that these patients are the ones with a more complex disease and a higher asthma morbidity (Anto et al., 2017b; Onder et al., 2015). This also suggests that these asthma patients may have an inadequate response to prescribed therapies since lung function was reduced and that they may represent a group of patients with the need of add-on

treatments, such as biological therapies; however, the choice of specific treatments will be more difficult considering the complexity introduced by having multiple phenotypes.

The more prevalent combinations of phenotypes observed in this study were AwObesity together with either B-Eos-high or B-Eos&FeNO-low phenotypes. This supports the view that obesity-related asthma, despite often suggested to be a separate asthma phenotype associated with non-eosinophilic airway inflammation, may also be associated with eosinophilic inflammation (Bates et al., 2017; Leiria, Martins, & Saad, 2015). Also, it is of note that there were 77 subjects with asthma that could not be classified as having any of the studied phenotypes, supporting the fact that there is a considerable number of asthma patients whose clinical phenotype is not easily classified (e.g. asthmatics with irreversible airflow obstruction, patients with similar airways symptoms but with different pattern of airway inflammation), which suggests the presence of sub-phenotypes (Agusti et al., 2016; Christenson et al., 2015; Froidure et al., 2016; Zedan, 2015).

Furthermore, in Study I, previous observations were extended regarding B-Eos and FeNO being independently associated with current asthma and asthma-related outcomes (Malinovschi et al., 2013). In this study, using the highest cut-off for FeNO (35 ppb), we still had more than 30% of asthma patients with B-Eos <300/mm^3, supporting the view that each biomarker partially reflects a different inflammatory pathway with a separate trigger mechanism and the use of both biomarkers seems to provide added value for asthma classification (Malinovschi et al., 2013; Malinovschi et al., 2016). A striking example is the treatment with type 2-high-driven therapies targeting Th2 cytokines (IL-5, and IL-4/IL-13) which have different acting pathways: the anti-IL-5 biologic Mepolizumab led to a decrease in blood and sputum eosinophils but not FeNO (Haldar et al., 2009; Pavord et al., 2012). This fact is in line with the hypothesis that FeNO is primarily driven by IL-4/IL-13. In contrast, treatment with the anti-IL-13 biologic Lebrikizumab led to an increase in blood eosinophils and a decrease in FeNO (Corren et al., 2011).

As shown by our research (Amaral et al., 2016; Amaral et al., 2018a), the complexity and unique features of the concomitant asthma phenotypes when categorizing asthma in adults, suggest that using only the “classical” (theory-driven) approach, do not have enough robustness and may require a broader interventional approach. A combination of hypothesis and data-driven approaches could result in better characterization of the asthma patients, as defended by other authors (Belgrave et al., 2016; Bousquet et al., 2016), including clinical information and several biomarkers. Also, using data-driven

(unsupervised) statistical methods could lead to iteratively progress our knowledge on asthma phenotypes, biomarkers, endotypes, and biology.

Different types of data-driven methods have been widely used in airway diseases, such as hierarchical (Moore et al., 2010), partitioning (Lefaudeux et al., 2017), and LCA (Couto et al., 2015; Amaral et al., 2018c). Notably, LCA appeared to account better for the heterogeneity of airways symptoms, compared to other commonly used data-driven approaches (e.g. PAM) (Amaral et al., 2018b), suggesting that LCA is a person-centred analysis as opposed to variable-centred analysis. Moreover, the application of the latent class assignments developed from a national data source has previously demonstrated higher degrees of generalizability (Evenson, Wen, Howard, & Herring, 2016).

However, as reported in Study II, using the same sample from Study I to derive new data-driven asthma phenotypes, the proportions of the hypothesis-driven phenotypes were similar between the two data-driven phenotypes obtained by LCA, when applying clinical and physiological variables commonly used to characterize asthma (Amaral et al., 2018a). This was the first study to compare previously defined hypothesis-driven asthma phenotypes with data-driven ones in a large population sample, representative of the US general population.

Overall, the data-driven phenotypes of current asthma only differed in symptom frequency and lung function parameters. The inflammatory biomarkers, presence of obesity, smoking status, age of asthma onset and self-reported hay fever were not different between classes. This fact suggests that, for the general asthma population, the clinical and physiological variables available to classify asthma and commonly used predefined cut-offs seem to be insufficient to identify specific phenotypes (Amaral et al., 2019d). However, using a less stringent asthma definition (ever asthma) and in subjects with poor clinical outcomes (difficult asthma), these variables were also suboptimal to differentiate asthma subgroups (Amaral et al., 2019c).

Previous studies using data-driven approaches contributed to the definition of clusters/phenotypes based on similarities in clinical and inflammatory biomarkers (Lefaudeux et al., 2017; Moore et al., 2010; Wu et al., 2014). However, these approaches have been mostly applied to patients with moderate to severe asthma and/or clinically based settings. Therefore, the generalization to the general asthma population may be limited.

Similarly, research efforts are being made to integrate clinical characteristics with available biomarkers to identify data-driven asthma phenotypes in children (Collins et

al., 2013; Depner et al., 2014). However, the obtained phenotypes vary on key features that are more pronounced during childhood, including natural history of wheeze over time (Henderson et al., 2008), suggesting that further research is required to compare data- and hypothesis-driven approaches to identify asthma phenotypes in children.

Previously, it has been reported that the same biomarkers have different discriminatory features when identifying data-driven asthma phenotypes. For example, Loza et al. (2016) found that, despite type 2 inflammation was a major characteristic for defining the phenotypes, it only distinguished two phenotypes. The other three identified were clinically distinguished by the degree of asthma control and reduced lung function. Similarly, in the Wu et al. (2014) study, variables for age of asthma onset, quality of life, symptoms, and medication use were the most discriminatory features, while the inflammatory profile was the less distinguishing feature of the asthma phenotypes.

Additionally, Study II highlighted that the fixed cut-offs values commonly used may potentially miss more complex, and yet unidentified phenotypes. Although the cut-off choices of the asthma phenotypes were based on studies that used the same FeNO and B-Eos cut-off for therapeutic decisions and showed better efficacy (Castro et al., 2015; Dweik et al., 2010; Wenzel et al., 2016), neither of them could differentiate asthma phenotypes, in an unsupervised manner. This fact suggests that these fixed cut-offs not only should be interpreted with consideration of the clinical context but should also be avoided. Furthermore, FeNO guidelines recognized that fixed cut-offs are weakly recommended based on their low quality of evidence (Dweik et al., 2011).

Likewise, blood eosinophilia has been used as a marker of eosinophilic asthma phenotype. However, the most accurate cut-off for blood eosinophilia has not yet been established. In fact, various studies have used different cut-offs, and so a single eosinophil count may not be enough to classify asthma phenotype with confidence (Casciano et al., 2016; Florence Schleich, Corhay, et al., 2016; Singh, Kolsum, Brightling, Locantore, Agusti, et al., 2014). Recent studies that use data-driven methods include absolute values/counts instead of using cut-offs (Hsiao, Lin, Wu, Wang, & Wang, 2019; Sendín-Hernández et al., 2018). Also, the use of reference equations for predicted values, adjusted to the individual characteristics of each subject, could potentially be a useful approach (Jacinto et al., 2018; Quanjer et al., 2012).

Thus, assessing the relevance of additional biomarkers that have been shown to be helpful in discriminating asthma phenotypes in population-based study settings, such as serum IgE (Patelis et al., 2012), combined with a broad set of variables from

different domains - clinical, physiologic, and/or disease features (such as objective assessment of atopy, nasal and ocular symptoms), and including them in the data-driven models, may result in the identification of more precise phenotypes, with the potential to be useful for clinicians and for population-based research.

Study III identified novel phenotypes of allergic respiratory disease (ARD), using more comprehensive disease features, such as using ocular symptoms as an independent disorder, and combining unsupervised and supervised analysis. This study revealed novel insights of ARD phenotyping, while helping confirm and integrate phenotypes previously reported.

Moreover, patient's profiles differed by their association with IgE, as two non-allergic ARD phenotypes (classes 1 and 5), three allergic phenotypes (classes 3, 4, and 6) and one with 50% participants being allergic (Class 2) were identified. Novel severe phenotypes of participants with co-occurrence of ocular, nasal and bronchial symptoms and exacerbation-prone (classes 5 and 6), were also revealed. Also, there was a class that was the mildest phenotype, without symptoms or inflammation (Class 1).

The obtained LCA-classes were derived from the general population and were very similar to previously published clinical phenotypes in non-allergic and allergic rhinoconjunctivitis patients (Di Lorenzo et al., 2011; Mølgaard et al., 2007), reinforcing the different patterns of multimorbidity in participants with rhinitis.

Class 5 comprised predominantly females, with low FeNO and B-Eos, and a high proportion of urgent healthcare use, suggesting that bronchial symptoms can be linked with NAR or rhinosinusitis (Shaaban et al., 2008). Moreover, participants in this class have some characteristics similar to the NAR phenotype obtained by cluster analysis in other studies (Burte et al., 2015; Kurukulaaratchy et al., 2015).

A polysensitized multimorbid phenotype was also identify in Study III, the Class 6, previously proposed by MeDALL (Mechanisms of the Development of ALLergy) (Anto et al., 2017; Bousquet et al., 2015) and now confirmed. This class was associated not only with rhinitis and asthma severity, but also to conjunctivitis. Furthermore, this class has some characteristics similar to those found in other unsupervised clustering studies, which has labelled as "late-onset, inflammation predominant" (de Vries et al., 2018; Haldar et al., 2008).

One surprising issue was the prevalence of conjunctivitis in the four classes. For the first time, ocular symptoms *per se* were used in LCA, being very informative, suggesting that they are essential to identify clusters of ARD patients. Also, it was

found that ocular symptoms were associated with the severity of nasal symptoms suggesting that rhinitis and rhinoconjunctivitis represent two distinct phenotypes. This is supported by another study that showed that the number of allergens recognized in the two phenotypes differ (low number in rhinitis, significantly higher number for rhinoconjunctivitis) (Siroux et al., 2019).

The unsupervised analysis did not identify clusters of participants having asthma, rhinitis, or conjunctivitis only, in this sample (Amaral et al., 2018c; Amaral et al., 2019a), suggesting that these conditions could be different manifestations of the same disease. Particularly, we found that there was no asthma cluster without rhinitis in agreement with the findings of the epidemiologic study European Community Respiratory Health Survey, where asthma alone represents less than 10% of the asthmatic population (Janson et al., 2001). Thus, it seems that in clinical practice, multimorbidity should always be investigated in ARD patients. This fact also holds for a Precision medicine approach based on treatable traits, rather than diagnostic labels, in the clinical management of the ARDs (Agusti et al., 2016; Bousquet, Vignola, & Demoly, 2003).

In Portugal, a previous study using an unsupervised clustering method was performed with a sample from a tertiary care outpatient clinic (Loureiro et al., 2015). Five clusters were obtained differing on age of disease onset, obesity, lung function, FeNO and disease severity. However, the lack of other biomarkers (e.g. serum IgE), nasal and ocular symptoms, and the fact that was performed in a different clinical setting, prevented the comparison to the LCA-phenotypes obtained in Study III.

To help classify patients in clinical settings and to distinguish between lower or higher degrees of airways allergic multimorbidity, a classification tree algorithm was generated by CART analysis using easily accessible parameters available in the clinical setting. CART is commonly used to develop reliable clinical decision rules in the development of new classification of patients into categories (Moore et al., 2010), because of their easily interpretive nature and ability to handle missing data (Marshall, 2001). In Study III, the tree algorithm reinforced the importance of including the presence of ocular symptoms in the expression of ARD phenotypes among other parameters.

These findings challenge the conventional disease classification of a “classical” clinical diagnosis organ-based approach to data-driven view, based on clinical characteristics and allergy profiles focusing on allergic multi-morbidities. Moreover, the concept of the importance of ocular symptoms in allergic multimorbidity was raised by a big data study using an App in uncharacterized users - MASK (MACVIA-ARIA Sentinel NetworK for

allergic rhinitis), and suggested by MeDALL (Bousquet et al., 2018). This hypothesis-generating study is now confirmed in a “classical” epidemiologic study. This is the first time in the field of allergy (and possibly other groups of diseases) that a big data analysis is confirmed by an epidemiologic study.

4.3 Strengths and limitations

Studies I and II, utilized data from the NHANES survey, a nationally representative survey of the civilian, non-institutionalized US population that uses a complex stratified, multistage probability sampling (CDC, 2017). Study III used data from the national and cross-sectional study conducted in the Portuguese general population, ICAR study (Sá-Sousa et al., 2019). Because of their cross-sectional design it was not possible to evaluate stability and interactions between phenotypes/classes longitudinally.

In Study I, it was also not possible to determine which phenotype occurred first in patients with concomitant phenotypes. Moreover, the “current asthma” and COPD definitions were based on self-reported diagnosis, rather than relying on lung function tests, so the acquired information is subject to recall and misclassification biases. However, in epidemiological studies, specific characteristics seem to be associated with better classification, depending on what questions we have available (Sá-Sousa et al., 2014), so we tried to use their combination to increase the acuity of the definition.

In the literature, we observed that not only current asthma was most frequently defined by the combination of: “ever diagnosed with asthma by a health professional” with “still having asthma” (Sá-Sousa et al., 2014); but also that in some studies, which validated asthma questionnaires with lung function tests and physician diagnosis, found that questions on wheeze are the most sensitive, while and questions such as “Have you ever had asthma?”, or questions on “waking with attacks of shortness of breath” and “morning tightness”, have higher specificity for asthma (De Marco, Cerveri, Bugiani, Ferrari, & Verlato, 1998; Toren, Brisman, & Jarvholm, 1993).

Similarly, the definition of COPD can be based on the diagnostic method (e.g. self-reported or spirometry), the criteria commonly used to define COPD (e.g. Global Initiative for Chronic Obstructive Lung Disease, ATS/ERS), and simultaneously the age group analysed (e.g. > 18 years or > 40 years). However, the most frequent

combination, using NHANES population, is a positive response to both “Has a doctor ever told you that you have chronic bronchitis?” and similar questions asked about “Has a doctor ever told you that you have emphysema?” (Diaz-Guzman, Khosravi, & Mannino, 2011; Tilert, Dillon, Paulose-Ram, Hnizdo, & Doney, 2013). Still, we opted to use a broader combination of questions to increase the accuracy of the definition: self-reported chronic bronchitis/emphysema + with age of diagnosis ≥ 40 years + having self-reported smoking history. Moreover, the choice of stratifying the analysis into less than 40 years-old or more than 40 years-old, was based on the Burden of Obstructive Lung Disease (BOLD) Initiative (Buist et al., 2005), that developed standardized methods for estimating COPD prevalence and for obtaining information about risk factors. The BOLD protocol included individuals aged 40 years and older as COPD develops over several decades of exposure to inhaled particulates (Buist et al., 2007; Vollmer et al., 2009) and this cut-off is frequently used in the literature (Lamprecht et al., 2011; Soriano et al., 2010; Toelle et al., 2013).

The label "asthma with concurrent COPD" (AwCOPD) was used as a proxy of ACOS, given that there is no standard definition of ACOS, and previous studies showing heterogeneity of ACOS by combining asthma and COPD diagnosis, i.e. by self-report of having current asthma and COPD (Mendy, Forno, Niyonsenga, Carnahan, & Gasana, 2018).

Moreover, given the high prevalence in the US population, in sample from the NHANES study, obesity is likely to be a comorbidity, rather than the primary reason for asthma (Gonzalez-Barcala et al., 2013); however, we defined the AwObesity phenotype as a separate group, since the interdependence on inflammatory markers to targeting different asthma therapies makes essential the accurate characterization of inflammation in obese asthmatic subjects (Amelink et al., 2013; Gibeon et al., 2013). Also, we performed an additional analysis excluding the AwObesity phenotype and the proportion of subjects with overlap of phenotypes remained high. Similar results were obtained if we do not consider only the B-Eos-high phenotype.

In Study I, the lack of other biomarkers in the NHANES years (2007-2012), prevented the analysis of other asthma phenotypes and the use of alternative definitions, such as the allergic asthma phenotypes based on total/specific IgE levels. This data was limited only to the 2005–2006 survey, however, data on FeNO was absent in this year range. Therefore, biomarkers of type-2 inflammation in both blood and exhaled air were analysed, available in the 2007-2012 NHANES survey years, and that has been shown previously to be independently related to asthma morbidity (Malinovschi et al., 2013,

2016). However, as there is no consensual definition of biomarker-defined asthma phenotypes, we based our definitions on cut-offs used in previous studies to discriminate patients in single asthma phenotypes (Castro et al., 2015; Dweik et al., 2010), rather than on any reported specificity or sensitivity for predicting asthma morbidity or response to therapeutics.

For high probability of airway inflammation, the cut-off value for FeNO has been suggested to be >50 ppb for adults (Loza et al., 2016). However, we chose a FeNO cut-off of 35 ppb, based on the mean baseline FeNO levels of patients included in randomized controlled trials of anti-IL-13 treatment (Corren et al., 2011; Dweik et al., 2010). Similarly, the variation of the cut-off of blood eosinophils counts is wide, even though they may help the stratification of type 2-high versus type 2-low asthma phenotypes (Froidure et al., 2016; Katial et al., 2017). As in our studies we have a non-severe, non-tertiary based population, we based our work on studies that either used the same cut-off for therapeutic decision or showed better efficacy in subjects with B-Eos≥ 300/mm^3 (Castro et al., 2015; Máspero, 2017).

In Study III, an important limitation is the decreased external validity, and therefore, less generalizability. This study, undertaken in Portugal, needs to be confirmed in other countries as regional variations exist (e.g. in allergen sensitizations). However, a study on multimorbidity and allergen sensitizations showed that similar data were observed in France and Sweden suggesting common biological mechanisms (Siroux et al., 2018).

The unsupervised analysis based on airways symptoms and allergy profile data remains a powerful approach toward ARD phenotyping (Bousquet et al., 2011; Bousquet et al., 2016; Garcia-Aymerich et al., 2015); however, a potential source of bias for unsupervised (data-driven) analysis is how individuals were included, because the interpretation of any unsupervised model dependent on how the participant selection is performed. However, latent classes provide a useful exploration tool for representing heterogeneity across the dimensions included in the model more than with traditional analytic approaches (e.g. regression analysis). Also described, was that the performance (i.e. the quality of estimation) of LCA is mainly affected by small sample sizes (Finch & Bronk, 2011) and few or/and poor-quality indicators/inputs (Marsh, Hau, Balla, & Grayson, 1998; Oertzen, Hertzog, Lindenberger, & Ghisletta, 2010), and we very strongly believe we have fulfilled these requirements.

Importantly, a major strength in Study III is that, not only the variables used in the LCA were obtained in the medical evaluation (being much more comprehensive than those used in the initial screening), but also variables regarding ocular symptoms that were

not part of the initial screening. Therefore, the literature and use of data analysts strongly supports the methods and interpretation of results we have chosen.

Additional variables could have been included in the model. Ideally, the data selected for input in the LCA model should comprehend all the health domains relevant to the understanding of the ARD to classify observations into discrete and mutually exclusive classes (Wang & Wang, 2012). However, a high number of variables potentially increases the risk of some dimensions or domains being allocated too much weight in the modelling process (Oberski, 2016; Pepe & Janes, 2007). We chose to use only clinical variables comprising the 4 main dimensions of the ARDs - ocular, nasal and bronchial symptoms, and allergic profiles.

While some previous studies used only unsupervised-clustering methods to identify phenotypes of ARD (Burte et al., 2015; Haldar et al., 2008; Kurukulaaratchy et al., 2015; Siroux et al., 2011), Study III extended this approach into providing information on the importance of the variables that best distinguish between the obtained classes, using CART analysis. To our knowledge, this is the first time CART analysis is studied in a population-based study of ARD. This analysis has various advantages over other methods, including multivariable logistic regression: it is a nonparametric method; results are summarized in a tree, much simpler to interpret and more practical in a clinical setting; and measures the variable relevance in the model (relative impact of the predictors on the output) (Sauve & Tuleau-Malot, 2014). Also, our tree algorithm based on parameters already used in clinical practice performed well using a training and a test set, suggesting a high potential to be further applied.

4.4 Future research

Our understanding of chronic diseases of the airways has evolved in the past decades. However, because of its heterogeneity and complexity, a considerable number of features of the disease must be considered when identifying or refining phenotypes of chronic diseases of the airways. Individual environmental factors, such as air pollution, tobacco smoke and/or indoor allergens, could also influence phenotypes (Gilmour et al., 2006; Ho, 2010; Lim et al., 2016; McCreanor et al., 2007). Further studies describing the overlap in patients with COPD, severe asthma and studies examining

asthma patients exposed to different environmental factors are needed, such as the comparison of subjects who live in cities versus rural areas.

The tree algorithm obtained in Study III, revealed a high potential to be further applied in assigning subjects into the six defined LCA-classes, using easily available variables in the clinical setting. Further studies to evaluate the clinical utility of this algorithm must be conducted in a separate and diverse population.

After this validation, future work should also address biologic associations, such as total/specific IgE, serum eosinophil cationic protein, and mechanisms within the phenotypes, to investigate treatment and outcome differences between them. Also, studies on the differentiation of COPD phenotypes, rhinitis phenotypes, allergic vs non-allergic conjunctivitis, and regarding symptoms' deterioration among non-smokers vs smokers/former-smokers and among different sensitizations, should be done, especially with prospective designs.

Future studies that combine comprehensive clinical, physiologic, and/or disease features with a broader availability of additional easily measurable biomarkers might provide several avenues for future research. Ideally, tailoring models to the individual should comprehend the optimal use of clustering methodologies and datasets incorporating genetic, epigenetic, and detailed molecular-level data (Wenzel, 2012; Pembrey et al., 2018).

Therefore, further improvements in the way that data are collected and stored to facilitate the standardization of data pre-processing and analyses should be developed and applied. Disease registries as data sources could be the solution to facilitate the gathering of data. These registries are considered as powerful tools to improve disease-related knowledge (Agency for Healthcare Research and Quality, 2014), such as the Portuguese Severe Asthma Registry, a national web-based disease registry of adult and paediatric severe asthma patients, that prospectively collects clinical data and is prepared for further data exchange (Sá-Sousa et al., 2018).

There are many potentialities regarding this longitudinal real-life data, particularly when combined disease registries with mobile health (mHealth) applications (e.g. MASK, MACVIA-ARIA Sentinel Network for the management of allergic rhinitis mobile (Bousquet, Schunemann, Fonseca, Samolinski, & Bachert, 2015), which are becoming increasingly popular among physicians, patients and the general public.

Moreover, variables collected longitudinally are of extreme importance to evaluate stability of the obtained phenotypes and further validate them. Clinically relevant

factors/components that could longitudinally predict persistent/new-onset respiratory symptoms remain to be further examined, particularly in subjects with asthma from general population. Furthermore, recent evidence indicates that there is plasticity among T-cell programming, influenced by a specific microenvironmental context (allergen exposure/atopy, genetic background, age, infection history, pollution, or diet) (Veldhoen et al., 2008; Lloyd & Saglani, 2013). However, a better understanding of the molecular mechanisms underlying T-cell plasticity in different asthma phenotypes are required perhaps with longitudinal analysis to map fluctuations in phenotypes and correlate with changes in clinical disease.

Data-driven methods have been increasingly used to derive asthma phenotypes; however, the resulting models should be interpreted with caution when translating the results into clinical practice. In the preliminary results of a systematic literature review, we have identified a significant heterogeneity of the variables and statistical methods used for phenotyping (Amaral et al., 2019b). As described in Appendix I, this work aims to systematically review asthma phenotypes derived with data-driven methods, using variables easily measurable in a clinical setting, and to summarize their consistency. To expand this preliminary work, we are currently analysing the differences between the data-driven asthma phenotypes of each study, how they differ according to the sample characteristics, data availability, variables used and applied methodology. First, using the clusters' labels (phenotypes) identified in the systematic review (Tables 1 and 2), relevant terms/keywords and phrases (words co-occurrence) will be automatically extracted with the unsupervised Rapid Automatic Keyword Extraction (RAKE) algorithm (Rose, Engel, Cramer & Cowley, 2010). Secondly, a multiple correspondence analysis (MCA) will be applied to the clusters' labels, aiming to detect and explore relationships between socio-demographic, clinical, functional, inflammation, atopy, and other variables, such as psychological/behaviour variables, asthma-related medication use and healthcare use, revealing patterns in a complex dataset. MCA allows the analysis of multivariate categorical data and visualization of the results in a graphical manner, to provide a global view of the data that is useful for interpretation (Greenacre, 1992; Sourial et al., 2010). The matrix data, comprising all asthma-related domain, will be converted into dimensions that are structured from the most explicative to the least and respective composite scores will be given. Finally, a hierarchical clustering will be performed with the scores from the MCA dimensions, using the Euclidian distance and Ward's clustering method, to identify the major groups of variables in an unsupervised manner. The result of agglomerative clustering will be visualized as a dendrogram, according with the order of the variables' similarity.

With these exploratory analyses we hope to provide further insights in how the multiple domains of asthma are related (between variables and between different categories/levels of each variable), taking into account the different patient's characteristics, derived from difference data sources (such as electronic health records, clinical registries, surveys, mHealth data).

5. Conclusion

The findings of this thesis bring novel insights into phenotyping chronic diseases of the airways by exploring multidimensional models supported by data-driven statistical approaches. This work challenges the conventional disease classification of a "classical" clinical diagnosis organ-based approach (theory-driven), to a combination of unsupervised analysis (data-driven) and clinical knowledge.

A prevalent overlap of commonly reported asthma phenotypes was observed among non-selected adults from the general population, with implications for objective asthma outcomes. The complexity and unique features of profiling chronic diseases of the airways required a broader data analysis approach, irrespective of using data-driven or theory-driven approaches. This was demonstrated when more comprehensive disease features were applied to a non-severe, non-tertiary based population, using the same data-driven method and it was revealed novel severe phenotypes of airways diseases.

This thesis suggests improvements to the way in which data-driven methods are used and combined with clinical information and biomarkers, and also provides practical advice and tools for development and validation of a population- and evidence-based, multidimensional model, based on each individual clinical, functional and immuno-inflammatory characteristics.

In summary, progress in the identification and refinement of phenotypes of chronic diseases of the airways requires embracing the Precision Medicine paradigm, going beyond simplistic concepts, and aiming to identify patient-to-patient variation. Given the current efforts to target the clinical and functional heterogeneity of patients' symptom profiles, conditions and responses to therapy, I hope that the adoption of ideas presented here can lead to improvements in the precision and completeness of the conclusions drawn from experimental data and contribute to future progress in this field.

6. References

Adeloye, D., Chua, S., Lee, C., Basquill, C., Papana, A., Theodoratou, E., … Rudan, I. (2015). Global and regional estimates of COPD prevalence: Systematic review and meta–analysis. *Journal of Global Health*, *5*(2), 1-17.

Agache, I., Akdis, C., Jutel, M., & Virchow, J. C. (2012). Untangling asthma phenotypes and endotypes. *Allergy*, *67*(7), 835–846.

Agache, I. O. (2013). From phenotypes to endotypes to asthma treatment. *Current Opinion in Allergy and Clinical Immunology,13*(3), 249-56.

Agache, I. O. (2014). Endotype Driven Treatment of Asthma. *Current Treatment Options in Allergy*, *1*(2), 198–212.

Agency for Healthcare Research and Quality. (2014). Registries for Evaluating Patient Outcomes: A User's Guide Volume 1. *AHRQ Publication, Vol 1*.

Agustí, A., Antó, J. M., Auffray, C., Barbé, F., Barreiro, E., Dorca, J., … Sznajder, J. I. (2015). Personalized Respiratory Medicine: Exploring the Horizon, Addressing the Issues. Summary of a BRN-AJRCCM Workshop Held in Barcelona on June 12, 2014. *American Journal of Respiratory and Critical Care Medicine*, *191*(4), 391–401.

Agustí, A., Bafadhel, M., Beasley, R., Bel, E. H., Faner, R., Gibson, P. G., … Pavord, I. D. (2017). Precision medicine in airway diseases: moving to clinical practice. *European Respiratory Journal, 50*(4), pii: 1701655.

Agusti, A., Bel, E., Thomas, M., Vogelmeier, C., Brusselle, G., Holgate, S., … Pavord, I. D. (2016). Treatable traits: toward precision medicine of chronic airway diseases. *European Respiratory Journal*, *47*(2), 410–419.

Agustí, A., Edwards, L. D., Rennard, S. I., MacNee, W., Tal-Singer, R., Miller, B. E., … Celli, B. (2012). Persistent Systemic Inflammation is Associated with Poor Clinical Outcomes in COPD: A Novel Phenotype. *PLoS ONE*, *7*(5), e37483.

Agusti, A., & MacNee, W. (2013). The COPD control panel: towards personalised medicine in COPD. *Thorax*, *68*(7), 687–690.

Akdis, C., Hellings, P. W., Agache, I., & EAACI. (2015). *Global Atlas of Allergic Rhinitis and Chronic Rhinosinusitis.* Switzerland: European Academy of Allergy and

Clinical Immunology.

Albert, P., Agusti, A., Edwards, L., Tal-Singer, R., Yates, J., Bakke, P., … Calverley, P. (2012). Bronchodilator responsiveness as a phenotypic characteristic of established chronic obstructive pulmonary disease. *Thorax*, *67*(8), 701–708.

Alves, R. de S. A., Vianna, F. de A. F., & Pereira, C. A. de C. (2008). Clinical phenotypes of severe asthma. *Jornal Brasileiro de Pneumologia*, *34*(9), 646–653.

Alving, K., & Malinovschi, A. (2010). Basic aspects of exhaled nitric oxide. In I. Horvath & J. C. Jongste (Eds.), *Exhaled Biomarkers* (pp. 1-31). European Respiratory Society Monograph, Vol. 49.

Amaral, R., Bousquet, J., Pereira, A.M., Araújo, L., Sá-Sousa, A., Jacinto, T… Fonseca, J.A. (2019a). Disentangling the heterogeneity of allergic respiratory diseases by latent class analysis reveals novel phenotypes". *Allergy, 74*(4), 698-708.

Amaral, R., Fonseca, J.A., Jacinto, T., Malinovschi, A., Janson, C., & Alving, K. (2016). High proportion of overlap between adult asthma phenotypes in a large population-based sample. *European Respiratory Journal,48*(Suppl. 60, PA568).

Amaral, R., Fonseca, J. A., Jacinto, T., Pereira, A. M., Malinovschi, A., Janson, C., & Alving, K. (2018a). Having concomitant asthma phenotypes is common and independently relates to poor lung function in NHANES 2007–2012. *Clinical and Translational Allergy*, *8*:13.

Amaral, R., Jacinto, T., Pereira, A., Almeida, R., & Fonseca, J. A. (2018b). A comparison of unsupervised methods based on dichotomous data to identify clusters of airways symptoms: latent class analysis and partitioning around medoids methods. *European Respiratory Journal, 52*(Suppl. 62, PA4429).

Amaral, R., Jacinto, T., Sousa-Pinto, B., & Fonseca, J. A. (2019b). Where do we stand with asthma phenotypes derived from data-driven methods? A systematic review. *Poster presentation at ERS International Congress 2019, Madrid, Spain*.

Amaral, R., Pereira, A. M., Jacinto, T., Malinovschi, A., Janson, C., Alving, K., & Fonseca, J.A. (2019c). Comparison of hypothesis- and data-driven asthma phenotypes in NHANES 2007-2012: the importance of comprehensive data availability. *Clinical and Translational Allergy,* 9:17.

Amaral, R., Pereira, A.M., Araújo, L., Sá-Sousa, A., Jacinto, T., Almeida, R. … Fonseca, J.A.. (2018c). Phenotyping allergic respiratory diseases: An

unsupervised classification using latent class analysis. *Allergy,73*(S105).

Amaral, R., Pereira, A. M., Jacinto, T., Malinovschi, A., Janson, C., Alving, K., & Fonseca, J.A. (2019d). Identification Of Asthma Phenotypes In The US General Population: A Latent Class Analysis Approach. *Oral communication and poster presentation at EAACI International Congress 2019, Lisbon, Portugal*.

Amelink, M., de Nijs, S. B., de Groot, J. C., van Tilburg, P. M. B., van Spiegel, P. I., Krouwels, F. H., … Bel, E. H. (2013). Three phenotypes of adult-onset asthma. *Allergy*, *68*(5), 674–680.

Anto, J. M., Bousquet, J., Akdis, M., Auffray, C., Keil, T., Momas, I., … Xu, C. J. (2017). Mechanisms of the Development of Allergy (MeDALL): Introducing novel concepts in allergy phenotypes. *Journal of Allergy and Clinical Immunology, 139*(2), 388-399.

Archer, K. J., Lemeshow, S., & Hosmer, D. W. (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis*, *51*(9), 4450–4464.

˜rron, J. R., Townsend, M. J., Keir, M. E., Yaspan, B. L., & Chan, A. C. (2015). Stratified medicine in inflammatory disorders: From theory to practice. *Clinical Immunology, 161*(1), 11-22.

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, *17*(9), 507–522.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2015). Big data in healthcare: Challenges and opportunities. In *2015 International Conference on Cloud Technologies and Applications (CloudTech)* (pp. 1–7). IEEE.

Bafadhel, M., McKenna, S., Terry, S., Mistry, V., Reid, C., Haldar, P., … Brightling, C. E. (2011). Acute Exacerbations of Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, *184*(6), 662–671.

Banfield, J. D., & Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, *49*(3), 803.

Baptist, A. P., Hao, W., Karamched, K. R., Kaur, B., Carpenter, L., & Song, P. X. K. (2018). Distinct Asthma Phenotypes Among Older Adults with Asthma. *The Journal of Allergy and Clinical Immunology: In Practice*, *6*(1), 244–249.e2.

Bárbara, C., Rodrigues, F., Dias, H., Cardoso, J., Almeida, J., Matos, M. J., ... Burney,

P. (2013). Chronic obstructive pulmonary disease prevalence in Lisbon, Portugal: the burden of obstructive lung disease study. *Revista Portuguesa de Pneumologia, 19*(3), 96-105

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., & Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, *380*(9836), 37–43.

Basile, A. O., & Ritchie, M. D. (2018). Informatics and machine learning to define the phenotype. *Expert Review of Molecular Diagnostics*, *18*(3), 219–226.

Bateman, E. D., Reddel, H. K., van Zyl-Smit, R. N., & Agusti, A. (2015). The asthma–COPD overlap syndrome: towards a revised taxonomy of chronic airways diseases? *The Lancet Respiratory Medicine*, *3*(9), 719–728.

Bates, J. H. T., Poynter, M. E., Frodella, C. M., Peters, U., Dixon, A. E., & Suratt, B. T. (2017). Pathophysiology to Phenotype in the Asthma of Obesity. *Annals of the American Thoracic Society*, *14*(Supplement_5), S395–S398.

Beaglehole, R., Ebrahim, S., Reddy, S., Voûte, J., & Leeder, S. (2007). Prevention of chronic diseases: a call to action. *The Lancet*, *370*(9605), 2152–2157.

Bednekoff, P. A. (2008). On Science and Statistics. *BioScience*, *58*(5), 381–381.

Belgrave, D., Henderson, J., Simpson, A., Buchan, I., Bishop, C., & Custovic, A. (2016). Disaggregating Asthma: Big Investigation vs. Big Data. *Journal of Allergy and Clinical Immunology, 139*(2), 400-407.

Bentayeb, M., Billionnet, C., Baiz, N., Derbez, M., Kirchner, S., & Annesi-Maesano, I. (2013). Higher prevalence of breathlessness in elderly exposed to indoor aldehydes and VOCs in a representative sample of French dwellings. *Respiratory Medicine, 107*(10), 1598-1607.

Berry, M. W., & Kogan, J. (2010). *Text Mining: Theory and Applications*. John Wiley & Sons, Inc.

Bielory, L. (2010). Allergic Conjunctivitis and the Impact of Allergic Rhinitis. *Current Allergy and Asthma Reports*, *10*(2), 122–134.

Bochenek, G., Kuschill-Dziurda, J., Szafraniec, K., Plutecka, H., Szczeklik, A., & Nizankowska-Mogilnicka, E. (2014). Certain subphenotypes of aspirin-exacerbated respiratory disease distinguished by latent class analysis. *Journal of Allergy and Clinical Immunology*, *133*(1), 98–103.e6.

Borish, L. (2016). The immunology of asthma. *Annals of Allergy, Asthma & Immunology*, *117*(2), 108–114.

Boudier, A., Curjuric, I., Basagana, X., Hazgui, H., Anto, J. M., Bousquet, J., … Siroux, V. (2013). 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*, *188*(5), 550–560.

Bousquet, J., Anto, J., Auffray, C., Akdis, M., Cambon-Thomsen, A., Keil, T., … Zuberbier, T. (2011a). MeDALL (Mechanisms of the Development of ALLergy): an integrated approach from phenotypes to systems medicine. *Allergy*, *66*(5), 596–604.

Bousquet, J., Anto, J. M., Akdis, M., Auffray, C., Keil, T., Momas, I., … Xu, C. (2016). Paving the way of systems biology and precision medicine in allergic diseases: the MeDALL success story. *Allergy*, *71*(11), 1513–1525.

Bousquet, J., Anto, J. M., Sterk, P. J., Adcock, I. M., Chung, K., Roca, J., … Auffray, C. (2011b). Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Medicine*, *3*(7), 43.

Bousquet, J., Anto, J. M., Wickman, M., Keil, T., Valenta, R., Haahtela, T., … von Hertzen, L. (2015). Are allergic multimorbidities and IgE polysensitization associated with the persistence or re-occurrence of foetal type 2 signalling? The MeDALL hypothesis. *Allergy*, *70*(9), 1062–1078.

Bousquet, J., Arnavielhe, S., Bedbrook, A., Fonseca, J., Morais Almeida, M., Todo Bom, A., … De Vries, G. (2018). The Allergic Rhinitis and its Impact on Asthma (ARIA) score of allergic rhinitis using mobile technology correlates with quality of life: The MASK study. *Allergy*, *73*(2), 505–510.

Bousquet, J., Dahl, R., & Khaltaev, N. (2007). Global alliance against chronic respiratory diseases. *Allergy*, *62*(3), 216–223.

Bousquet, J., Devillier, P., Anto, J. M., Bewick, M., Haahtela, T., Arnavielhe, S., … Laune, D. (2018). Daily allergic multimorbidity in rhinitis using mobile technology: A novel concept of the MASK study. *Allergy*, *73*(8), 1622–1631.

Bousquet, J., Heinzerling, L., Bachert, C., Papadopoulos, N. G., Bousquet, P. J., Burney, P. G., … Demoly, P. (2012). Practical guide to skin prick tests in allergy to aeroallergens. *Allergy*, *67*(1), 18–24.

Bousquet, J., Khaltaev, N., Cruz, A. A., Denburg, J., Fokkens, W. J., Togias, A., …

Williams, D. (2008). Allergic Rhinitis and its Impact on Asthma (ARIA) 2008*. *Allergy*, *63*(SUPPL. 86), 8–160.

Bousquet, J., Schunemann, H. J., Fonseca, J., Samolinski, B., & Bachert, C. (2015). MACVIA-ARIA Sentinel NetworK for allergic rhinitis (MASK-rhinitis): The new generation guideline implementation. *Allergy*, *70*, 1372–1392.

Bousquet, J., Vignola, A. M., & Demoly, P. (2003). Links between rhinitis and asthma. *Allergy*, *58*(8), 691–706.

Bousquet, P. J., Combescure, C., Klossek, J. M., Daurès, J. P., & Bousquet, J. (2009). Change in visual analog scale score in a pragmatic randomized cluster trial of allergic rhinitis. *Journal of Allergy and Clinical Immunology*, *123*(6), 1349–1354.

Brightling, C. E., Monteiro, W., Ward, R., Parker, D., Morgan, M. D. L., Wardlaw, A. J., & Pavord, I. D. (2000). Sputum eosinophilia and short-term response to prednisolone in chronic obstructive pulmonary disease: A randomised controlled trial. *The Lancet, 356*(9240), 1480-1485.

Brinkman, P., Wagener, A. H., Hekking, P. P., Bansal, A. T., Maitland-van der Zee, A. H., Wang, Y., … Sterk, P. J. (2018). Identification and prospective stability of electronic nose (eNose)-derived inflammatory phenotypes in patients with severe asthma. *Journal of Allergy and Clinical Immunology, pii: S0091-6749*(18), 31722-6.

Buist, A. S., McBurnie, M. A., Vollmer, W. M., Gillespie, S., Burney, P., Mannino, D. M., … Nizankowska-Mogilnicka, E. (2007). International variation in the prevalence of COPD (The BOLD Study): a population-based prevalence study. *The Lancet*, *370*(9589), 741–750.

Buist, A. S., Vollmer, W. M., Sullivan, S. D., Weiss, K. B., Lee, T. A., Menezes, A. M. B., … Burney, P. G. J. (2005). The Burden of Obstructive Lung Disease Initiative (BOLD): Rationale and design. *COPD: Journal of Chronic Obstructive Pulmonary Disease, 2*(2), 277-283.

Burgel, P. R., Paillasseur, J. L., & Roche, N. (2014). Identification of Clinical Phenotypes Using Cluster Analyses in COPD Patients with Multiple Comorbidities. *BioMed Research International*, *2014*, 1–9.

Burte, E., Bousquet, J., Varraso, R., Gormand, F., Just, J., Matran, R., … Nadif, R. (2015). Characterization of Rhinitis According to the Asthma Status in Adults Using an Unsupervised Approach in the EGEA Study. *PLoS One*, *10*(8),

e0136191.

Caillez, F., & Kuntz, P. (1996). A contribution to the study of the metric and Euclidean structures of dissimilarities. *Psychometrika*, *61*(2), 241–253.

Carr, T. F., Zeki, A. A., & Kraft, M. (2018). Eosinophilic and noneosinophilic asthma. *American Journal of Respiratory and Critical Care Medicine, 97*(1), 22-37.

Carreiro-Martins, P., Gomes-Belo, J., Papoila, A. L., Caires, I., Palmeiro, T., Gaspar-Marques, J., … Neuparth, N. (2016). Chronic respiratory diseases and quality of life in elderly nursing home residents. *Chronic Respiratory Disease*, *13*(3), 211–219.

Casciano, J., Krishnan, J. A., Small, M. B., Buck, P. O., Gopalan, G., Li, C., … Dotiwala, Z. (2016). Value of peripheral blood eosinophil markers to predict severity of asthma. *BMC Pulmonary Medicine*, *16*(1), 109.

Castro, M., Zangrilli, J., Wechsler, M. E., Bateman, E. D., Brusselle, G. G., Bardin, P., … Korn, S. (2015). Reslizumab for inadequately controlled asthma with elevated blood eosinophil counts: results from two multicentre, parallel, double-blind, randomised, placebo-controlled, phase 3 trials. *The Lancet Respiratory Medicine*, *3*(5), 355–366.

CDC. (2011a). National Health and Nutrition Examination Survey. Respiratory Health ENO Procedures Manual. Retrieved from www.cdc.gov/nchs/data/nhanes/nhanes_11_12/Respiratory_Health_ENO_Procedures_Manual.pdf

CDC. (2011b). Respiratory Health Spirometry Procedures Manual. Retrieved from https://www.cdc.gov/%0Anchs/data/nhanes/nhanes_11_12/https://www.cdc.gov/nchs/data/nhanes/nhanes_11_12/spirometry_procedures_manual.pdf

CDC. (2013). National Health and Nutrition Examination 2011-2012. Complete Blood Count. Retrieved from http://www.cdc.gov/nchs/data/nhanes/nhanes_11_12/cbc_g_met_he.pdf

CDC. (2014). National Health and Nutrition Examination Survey 1988-2012. Prescription Medications - Drug Information. Retrieved from http://wwwn.cdc.gov/Nchs/Nhanes/1999-2000/RXQ_DRUG.htm

CDC. (2017). About the National Health and Nutrition Examination Survey. Retrieved from http://www.cdc.gov/nchs/nhanes/about_nhanes.htm

CDC & National Center for Health Statistics. (2013). Specifying weightning parameters. Retrieved August 20, 2011, from https://www.cdc.gov/nchs/tutorials/nhanes/SurveyDesign/Weighting/intro.htm

Cecere, L. M., Slatore, C. G., Uman, J. E., Evans, L. E., Udris, E. M., Bryson, C. L., & Au, D. H. (2012). Adherence to Long-Acting Inhaled Therapies among Patients with Chronic Obstructive Pulmonary Disease (COPD). *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *9*(3), 251–258.

Celli, B. R., & Agustí, A. (2018). COPD: time to improve its taxonomy? *ERJ Open Research*, *4*(1), 00132-2017.

Chanez, P., Wenzel, S. E., Anderson, G. P., Anto, J. M., Bel, E. H., Boulet, L. P., … Vachier, I. (2007). Severe asthma in adults: What are the important questions? *Journal of Allergy and Clinical Immunology, 19*(6), 1337-1348.

Chapman, K. R., Burdon, J. G. W., Piitulainen, E., Sandhaus, R. A., Seersholm, N., Stocks, J. M., … McElvaney, N. G. (2015). Intravenous augmentation treatment and lung density in severe α1 antitrypsin deficiency (RAPID): a randomised, double-blind, placebo-controlled trial. *The Lancet*, *386*(9991), 360–368.

Christenson, S. A., Steiling, K., van den Berge, M., Hijazi, K., Hiemstra, P. S., Postma, D. S., … Woodruff, P. G. (2015). Asthma–COPD Overlap. Clinical Relevance of Genomic Signatures of Type 2 Inflammation in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, *191*(7), 758–766.

Chu, W. W. (Ed.). (2014). *Data Mining and Knowledge Discovery for Big Data* (Vol. 1). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chung, K. F., Wenzel, S. E., Brozek, J. L., Bush, A., Castro, M., Sterk, P. J., … Teague, W. G. (2014). International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *European Respiratory Journal*, *43*(2), 343–373.

Cibella, F., Ferrante, G., Cuttitta, G., Bucchieri, S., Melis, M. R., La Grutta, S., & Viegi, G. (2015). The Burden of Rhinitis and Rhinoconjunctivitis in Adolescents. *Allergy, Asthma & Immunology Research*, *7*(1), 44.

Collins, L. M., & Lanza, S. T. (2009). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Collins, S. A., Pike, K. C., Inskip, H. M., Godfrey, K. M., Roberts, G., Holloway, J. W., ...

Southampton Women's Survey Study, G. (2013). Validation of novel wheeze phenotypes using longitudinal airway function and atopic sensitization data in the first 6 years of life: evidence from the Southampton Women's survey. *Pediatric Pulmonology, 48*(7), 683-692.

Corren, J., Lemanske, R. F., Hanania, N. A., Korenblat, P. E., Parsey, M. V, Arron, J. R., … Matthews, J. G. (2011). Lebrikizumab Treatment in Adults with Asthma. *New England Journal of Medicine*, *365*(12), 1088–1098.

Couto, M., Stang, J., Horta, L., Stensrud, T., Severo, M., Mowinckel, P., … Carlsen, K. H. (2015). Two distinct phenotypes of asthma in elite athletes identified by latent class analysis. *Journal of Asthma*, *52*(9), 897–904.

Cruz, A. A., Popov, T., Pawankar, R., Annesi-Maesano, I., Fokkens, W., Kemp, J., … Bousquet, J. (2007). Common characteristics of upper and lower airways in rhinitis and asthma: ARIA update, in collaboration with GA 2 LEN. *Allergy*, *62*(s84), 1–41.

De Greve, G., Hellings, P. W., Fokkens, W. J., Pugin, B., Steelant, B., & Seys, S. F. (2017). Endotype-driven treatment in chronic upper airway diseases. *Clinical and Translational Allergy*, *7*(1), 22.

De Marco, R., Cerveri, I., Bugiani, M., Ferrari, M., & Verlato, G. (1998). An undetected burden of asthma in Italy: The relationship between clinical and epidemiological diagnosis of asthma. *European Respiratory Journal*, *11*(3), 599–605.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122–135.

De Vries, R., Dagelet, Y. W. F., Spoor, P., Snoey, E., Jak, P. M. C., Brinkman, P., … Sterk, P. J. (2018). Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. *European Respiratory Journal*, *51*(1), 1701817.

Deccache, A., Didier, A., Mayran, P., Jeziorski, A., & Raherison, C. (2018). Adapter l'accompagnement thérapeutique de l'asthmatique en fonction des profils médico-psycho-sociaux. Enseignements de l'enquête REALISE™ chez des patients asthmatiques adultes et repères éducationnels pour la pratique. *Revue Des Maladies Respiratoires*, *35*(3), 313–323.

Delude, C. M. (2015). Deep phenotyping: The details of disease. *Nature*, *527*(7576), S14–S15.

Delgado-Eckert, E., Fuchs, O., Kumar, N., Pekkanen, J., Dalphin, J. C., Riedler, J., …

Frey, U. (2018). Functional phenotypes determined by fluctuation-based clustering of lung function measurements in healthy and asthmatic cohort participants. *Thorax*, *73*(2), 107–115.

Demoly, P., & Bousquet, P. J. (2008). Editorial: Links between allergic rhinitis and asthma still reinforced. *Allergy*, *63*(3), 251–254.

Depner, M., Fuchs, O., Genuneit, J., Karvonen, A. M., Hyvarinen, A., Kaulek, V., ... Pasture Study Group (2014). Clinical and epidemiologic phenotypes of childhood asthma. *American Journal of Respiratory and Critical Care Medicine, 189*(2), 129-138.

Di Lorenzo, G., Pacor, M. L., Amodio, E., Leto-Barone, M. S., La Piana, S., D'Alcamo, A., … Di Bona, D. (2011). Differences and Similarities between Allergic and Nonallergic Rhinitis in a Large Sample of Adult Patients with Rhinitis Symptoms. *International Archives of Allergy and Immunology*, *155*(3), 263–270.

Diaz-Guzman, E., Khosravi, M., & Mannino, D. M. (2011). Asthma, Chronic Obstructive Pulmonary Disease, and Mortality in the U.S. Population. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *8*(6), 400–407.

Donnelly, L.E. (2010). Exhaled breath condensate: Nitric oxide-related compounds. In I. Horvath & J. C. Jongste (Eds.), *Exhaled Biomarkers* (pp. 207–216). European Respiratory Society Monograph.

Droessaert, V., Timmermans, M., Dekimpe, E., Seys, S., Ceuppens, J. L., Fokkens, W. J., & Hellings, P. W. (2016). Real-life study showing better control of allergic rhinitis by immunotherapy than regular pharmacotherapy. *Rhinology, 54*(3), 214-220.

Dweik, R. A., Boggs, P. B., Erzurum, S. C., Irvin, C. G., Leigh, M. W., Lundberg, J. O., … Taylor, D. R. (2011). An official ATS clinical practice guideline: Interpretation of exhaled nitric oxide levels (FENO) for clinical applications. *American Journal of Respiratory and Critical Care Medicine*, *184*(5), 602–615.

Dweik, R. A., Sorkness, R. L., Wenzel, S., Hammel, J., Curran-Everett, D., Comhair, S. A. A., … Erzurum, S. C. (2010). Use of Exhaled Nitric Oxide Measurement to Identify a Reactive, at-Risk Phenotype among Patients with Asthma. *American Journal of Respiratory and Critical Care Medicine*, *181*(10), 1033–1041.

Evenson, K. R., Wen, F., Howard, A. G., & Herring, A. H. (2016). Applying latent class assignments for accelerometry data to external populations: Data from the

National Health and Nutrition Examination Survey 2003–2006. *Data in Brief*, *9*, 926–930.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. *Cluster Analysis: Fifth Edition*. Chichester, UK: John Wiley & Sons, Ltd.

Fahy, J. V. (2015). Type 2 inflammation in asthma — present in most, absent in many. *Nature Reviews Immunology*, *15*(1), 57–65.

Finch, W. H., & Bronk, K. C. (2011). Conducting Confirmatory Latent Class Analysis Using M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(1), 132–151.

Fingleton, J., Huang, K., Weatherall, M., Guo, Y., Ivanov, S., Bruijnzeel, P., … Wang, C. (2017). Phenotypes of symptomatic airways disease in China and New Zealand. *European Respiratory Journal*, *50*(6).

Fingleton, J., Travers, J., Williams, M., Charles, T., Bowles, D., Strik, R., … Beasley, R. (2015). Treatment responsiveness of phenotypes of symptomatic airways obstruction in adults. *Journal of Allergy and Clinical Immunology*, *136*(3), 601–609.

Flegal, K., Carroll, M., Kuczmarski, R., & Johnson, C. (1998). Overweight and obesity in the United States: prevalence and trends, 1960–1994. *International Journal of Obesity*, *22*(1), 39–47.

Flores, M., Glusman, G., Brogaard, K., Price, N. D., & Hood, L. (2013). P4 medicine: how systems medicine will transform the healthcare sector and society. *Personalized Medicine*, *10*(6), 565–576.

Fonseca, J. A., Nogueira-Silva, L., Morais-Almeida, M., Azevedo, L., Sa-Sousa, A., Branco-Ferreira, M., … Bousquet, J. (2010). Validation of a questionnaire (CARAT10) to assess rhinitis and asthma in patients with asthma. *Allergy*, *65*(8), 1042–1048.

Fonseca-Antunes, A., Bárbara, C., Melo-Gomes, E. (2013). Portuguese National Programme for Respiratory Diseases/2012-2016. Lisboa, Portugal: Ministério da Saúde - Direção Geral de Saúde de Portugal.

Forum of International Respiratory Societies (2017). *The Global Impact of Respiratory Disease* (2nd ed.). Sheffield, European Respiratory Society.

Froidure, A., Mouthuy, J., Durham, S. R., Chanez, P., Sibille, Y., & Pilette, C. (2016).

Asthma phenotypes and IgE responses. *European Respiratory Journal*, *47*, 304–319.

Galas, D. J., & Hood, L. (2009). Systems Biology and Emerging Technologies Will Catalyze the Transition from Reactive Medicine to Predictive, Personalized, Preventive and Participatory (P4) Medicine. *Interdisciplinary Bio Central, 1*(2), 1-4

Garcia-Aymerich, J., Benet, M., Saeys, Y., Pinart, M., Basagaña, X., Smit, H. A., … Antó, J. M. (2015). Phenotyping asthma, rhinitis and eczema in MeDALL population-based birth cohorts: an allergic comorbidity cluster. *Allergy*, *70*(8), 973–984.

Ghebre, M. A., Bafadhel, M., Desai, D., Cohen, S. E., Newbold, P., Rapley, L., … Brightling, C. E. (2015). Biological clustering supports both “dutch” and “british” hypotheses of asthma and chronic obstructive pulmonary disease. *Journal of Allergy and Clinical Immunology, 135*(1):63-72.

Gibeon, D., Batuwita, K., Osmond, M., Heaney, L. G., Brightling, C. E., Niven, R., … Menzies-Gow, A. (2013). Obesity-associated severe asthma represents a distinct clinical phenotype: analysis of the British Thoracic Society Difficult Asthma Registry Patient cohort according to BMI. *Chest*, *143*(2), 406–414.

Gibson, G. J., Loddenkemper, R., Sibille, Y. & Lundbäck, B. (2013). *The European Lung White Book: Respiratory Health and Disease in Europe* (2nd ed). European Respiratory Society, Sheffield.

Gibson, P. G., & Simpson, J. L. (2009). The overlap syndrome of asthma and COPD: what are its features and how important is it? *Thorax*, *64*(8), 728–735.

Gilmour, M. I., Jaakkola, M. S., London, S. J., Nel, A. E., & Rogers, C. A. (2006). How exposure to environmental tobacco smoke, outdoor air pollutants, and increased pollen burdens influences the incidence of asthma*. Environmental Health Perspectives, 114*(4), 627-633.

Global Initiative for Asthma. (2018). Global Strategy for Asthma Management and Prevention. Global Initiative for Asthma. Retrieved from http://www.ginasthma.org.

Gligorijević, V., Malod-Dognin, N., & Pržulj, N. (2016). Integrative methods for analyzing big data in precision medicine. *Proteomics, 16*(5), 741-758.

Gonzalez-Barcala, F. J., Pertega, S., Perez-Castro, T., Sampedro, M., Sanchez-Lastres, J., San-Jose-Gonzalez, M. A., … Lopez-Silvarrey, A. (2013). Obesity and asthma: An association modified by age. *Allergologia et Immunopathologia*, *41*(3),

176–180.

Goodman, L. A. (2009). Latent Class Analysis: The Empirical Study of Latent Types, Latent Variables, and Latent Structures. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 3–55). Cambridge: Cambridge University Press.

Green, M. A., Strong, M., Razak, F., Subramanian, S. V., Relton, C., & Bissell, P. (2016). Who are the obese? A cluster analysis exploring subgroups of the obese. *Journal of Public Health*, *38*(2), 258–264.

Greenacre M. (1992). Correspondence analysis in medical research. *Statistical Methods in Medical Research, 1*(1), 97–117.

Greiner, A. N., Hellings, P. W., Rotiroti, G., & Scadding, G. K. (2011). Allergic rhinitis. *The Lancet*, *378*(9809), 2112–2122.

Hahtola, P. A., & Hahtola, K. H. (2003). Asthma in light of patient documents. A factor analytic study. *Annals of the Academy of Medicine, Singapore*, *32*(3), 392–402.

Haldar, P., Brightling, C. E., Hargadon, B., Gupta, S., Monteiro, W., Sousa, A., … Pavord, I. D. (2009). Mepolizumab and Exacerbations of Refractory Eosinophilic Asthma. *New England Journal of Medicine*, *360*(10), 973–984.

Haldar, P., Pavord, I. D., Shaw, D. E., Berry, M. A., Thomas, M., Brightling, C. E., … Green, R. H. (2008). Cluster analysis and clinical asthma phenotypes. *American Journal of Respiratory and Critical Care Medicine*, *178*(3), 218–224.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Boston: Morgan Kaufmann.

Han, M. K., Agusti, A., Calverley, P. M., Celli, B. R., Criner, G., Curtis, J. L., … Martinez, F. J. (2010). Chronic Obstructive Pulmonary Disease Phenotypes. *American Journal of Respiratory and Critical Care Medicine*, *182*(5), 598–604.

Hankinson, J. L., Kawut, S. M., Shahar, E., Smith, L. J., Stukovsky, K. H., & Barr, R. G. (2010). Performance of American Thoracic Society-Recommended Spirometry Reference Values in a Multiethnic Sample of Adults: the multi-ethnic study of atherosclerosis (MESA) lung study. *Chest, 137*(1), 138-145.

Hankinson, J. L., Odencrantz, J. R., & Fedan, K. B. (1999). Spirometric Reference Values from a Sample of the General U.S. Population. *American Journal of Respiratory and Critical Care Medicine*, *159*(1), 179–187.

Hansen, G., Berry, G., DeKruyff, R. H., & Umetsu, D. T. (1999). Allergen-specific Th1 cells fail to counterbalance Th2 cell–induced airway hyperreactivity but cause severe airway inflammation. *Journal of Clinical Investigation*, *103*(2), 175–183.

Hastan, D., Fokkens, W. J., Bachert, C., Newson, R. B., Bislimovska, J., Bockelbrink, A., … Burney, P. (2011). Chronic rhinosinusitis in Europe - An underestimated disease. A GA(2)LEN study. *Allergy, 66*(9), 1216-1223.

Hastie, A. T., Moore, W. C., Li, H., Rector, B. M., Ortega, V. E., Pascual, R. M., … Bleecker, E. R. (2013). Biomarker surrogates do not accurately predict sputum eosinophil and neutrophil percentages in asthmatic subjects. *Journal of Allergy and Clinical Immunology*, *132*(1), 72–80.e12.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. *Book*. New York, NY: Springer New York.

Heaney, L. G., & McGarvey, L. P. A. (2017). Personalised Medicine for Asthma and Chronic Obstructive Pulmonary Disease. *Respiration*, *93*(3), 153–161.

Heinzerling, L., Frew, A. J., Bindslev-Jensen, C., Bonini, S., Bousquet, J., Bresciani, M., … Zuberbier, T. (2005). Standard skin prick testing and sensitization to inhalant allergens across Europe - a survey from the GA2LEN network. *Allergy*, *60*(10), 1287–1300.

Heinzerling, L., Mari, A., Bergmann, K.-C., Bresciani, M., Burbach, G., Darsow, U., … Lockey, R. (2013). The skin prick test – European standards. *Clinical and Translational Allergy*, *3*(1), 3.

Hekking, P. P. W., & Bel, E. H. (2014). Developing and Emerging Clinical Asthma Phenotypes. *The Journal of Allergy and Clinical Immunology: In Practice*, *2*(6), 671–680.

Hellings, P. W., Fokkens, W. J., Bachert, C., Akdis, C. A., Bieber, T., Agache, I., … Groups, A. E. W. (2017). Positioning the principles of precision medicine in care pathways for allergic rhinitis and chronic rhinosinusitis - A EUFOREA-ARIA-EPOS-AIRWAYS ICP statement. *Allergy*, *72*(9 PG-1297-1305), 1297–1305.

Hellings, P. W., Klimek, L., Cingi, C., Agache, I., Akdis, C., Bachert, C., … Fokkens, W. J. (2017). Non-allergic rhinitis: Position paper of the European Academy of Allergy and Clinical Immunology. *Allergy*, *72*(11), 1657–1665.

Henderson, J., Granell, R., Heron, J., Sherriff, A., Simpson, A., Woodcock, A., ... Sterne, J. A. (2008). Associations of wheezing phenotypes in the first 6 years of

life with atopy, lung function and airway responsiveness in mid-childhood. *Thorax, 63*(11), 974-980.

Henry, R. L. (2002). Invasive monitoring of airway inflammation. *The Medical Journal of Australia*, 177 Suppl, S57-58.

Hirano, T., & Matsunaga, K. (2018). Late-onset asthma: current perspectives. *Journal of Asthma and Allergy*, *11*, 19–27.

Ho, S. M. (2010). Environmental epigenetics of asthma: an update. *Journal of Allergy and Clinical Immunology, 126*(3), 453-465.

Hodson, R. (2016). Precision medicine. *Nature*, *537*(7619), S49.

Hsiao, H. P., Lin, M. C., Wu, C. C., Wang, C. C., & Wang, T. N. (2019). Sex-Specific Asthma Phenotypes, Inflammatory Patterns, and Asthma Control in a Cluster Analysis. *The Journal of Allergy and Clinical Immunology: In Practice*, *7*(2), 556–567.e15.

Huddleston, S. H., & Brown, G. G. (2018). Machine Learning. In *INFORMS Analytics Body of Knowledge* (pp. 231–274). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Hüls, A., Abramson, M. J., Sugiri, D., Fuks, K., Krämer, U., Krutmann, J., & Schikowski, T. (2019). Nonatopic eczema in elderly women: Effect of air pollution and genes. *Journal of Allergy and Clinical Immunology*, *143*(1 PG-378-385.e9), 378–385.e9.

Ibrahim, B., Basanta, M., Cadden, P., Singh, D., Douce, D., Woodcock, A., & Fowler, S. J. (2011). Non-invasive phenotyping using exhaled volatile organic compounds in asthma. *Thorax, 66*(9), 804-809.

Ilmarinen, P., Tuomisto, L. E., Niemelä, O., Tommola, M., Haanpää, J., & Kankaanranta, H. (2017). Cluster analysis on longitudinal data of patients with adult-onset asthma. *The Journal of Allergy and Clinical Immunology: In Practice*, *5*(4), 967–978.e3.

Jacinto, T., Amaral, R., Malinovschi, A., Janson, C., Fonseca, J., & Alving, K. (2018). Exhaled NO reference limits in a large population-based sample using the Lambda-Mu-Sigma method. *Journal of Applied Physiology*, *125*(5), 1620–1626.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

James, A. L., Palmer, L. J., Kicic, E., Maxwell, P. S., Lagan, S. E., Ryan, G. F., & Musk, A. W. (2005). Decline in Lung Function in the Busselton Health Study.

American Journal of Respiratory and Critical Care Medicine, *171*(2), 109–114.

Jang, A. S., Kwon, H.-S., Cho, Y. S., Bae, Y. J., Kim, T. B., Park, J. S., … Park, C. S. (2013). Identification of Subtypes of Refractory Asthma in Korean Patients by Cluster Analysis. *Lung*, *191*(1), 87–93.

Janson, C., Anto, J., Burney, P., Chinn, S., de Marco, R., Heinrich, J., … Wjst, M. (2001). The European Community Respiratory Health Survey: what are the main results so far? *European Respiratory Journal*, *18*(3), 598–611.

Janssens, T., Verleden, G., & Van Den Bergh, O. (2012). Symptoms, lung function, and perception of asthma control: An exploration into the heterogeneity of the asthma control construct. *Journal of Asthma*, *49*(1), 63–69.

Jeong, A., Imboden, M., Hansen, S., Zemp, E., Bridevaux, P.-O., Lovison, G., … Probst-Hensch, N. (2017). Heterogeneity of obesity-asthma association disentangled by latent class analysis, the SAPALDIA cohort. *Respiratory Medicine*, *125*, 25–32.

Jia, G., Erickson, R. W., Choy, D. F., Mosesova, S., Wu, L. C., Solberg, O. D., … Arron, J. R. (2012). Periostin is a systemic biomarker of eosinophilic airway inflammation in asthmatic patients. *Journal of Allergy and Clinical Immunology*, *130*(3), 647–654.e10.

Katial, R. K., Bensch, G. W., Busse, W. W., Chipps, B. E., Denson, J. L., Gerber, A. N., … Wechsler, M. E. (2017). Changing Paradigms in the Treatment of Severe Asthma: The Role of Biologic Therapies. *The Journal of Allergy and Clinical Immunology: In Practice*, *5*(2), S1–S14.

Kaufman, L., & Rousseeuw, P. (1990). Partitioning Around Medoids (Program PAM). In L. Kaufman & P. Rousseeuw (Eds.), *Finding Groups in Data: An Introduction to Cluster Analysis* (pp. 68–125). Hoboken: John Wiley & Sons, Inc.

Khatry, D. B., Gossage, D. L., Geba, G. P., Parker, J. M., Jarjour, N. N., Busse, W. W., & Molfino, N. A. (2015). Discriminating sputum-eosinophilic asthma: Accuracy of cutoffs in blood eosinophil measurements versus a composite index, ELEN. *Journal of Allergy and Clinical Immunology*, *136*(3), 812–814.e2.

Khusial, R. J., Sont, J. K., Loijmans, R. J. B., Snoeck-Stroband, J. B., Assendelft, P. J. J., Schermer, T. R. J., & Honkoop, P. J. (2017). Longitudinal outcomes of different asthma phenotypes in primary care, an observational study. *Npj Primary Care Respiratory Medicine*, *27*(1), 55.

Kim, M.A., Shin, S.W., Park, J.S., Uh, S.T., Chang, H. S., Bae, D.J., … Park, C.S. (2017). Clinical Characteristics of Exacerbation-Prone Adult Asthmatics Identified by Cluster Analysis. *Allergy, Asthma & Immunology Research*, *9*(6), 483.

Kim, M. H., Rhee, C. K., Kim, K., Kim, S. H., Lee, J. Y., Kim, Y. H., … Lee, J. H. (2018). Heterogeneity of asthma and COPD overlap. *International Journal of Chronic Obstructive Pulmonary Disease*, *13*, 1251–1260.

Kim, T.B., Jang, A.S., Kwon, H.S., Park, J.S., Chang, Y.S., Cho, S.H., … Park, C.S. (2013). Identification of asthma clusters in two independent Korean adult asthma cohorts. *European Respiratory Journal*, *41*(6), 1308–1314.

Kobayashi, S., Hanagama, M., Yamanda, S., Ishida, M., & Yanai, M. (2016). Inflammatory biomarkers in asthma-COPD overlap syndrome. *International Journal of Chronic Obstructive Pulmonary Disease*, *11*, 2117–2123.

Koch, G., & Landis, J. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.

König, I. R., Fuchs, O., Hansen, G., von Mutius, E., & Kopp, M. V. (2017). What is precision medicine? *European Respiratory Journal*, *50*(4), pii: 1700391.

Konno, S., Taniguchi, N., Makita, H., Nakamaru, Y., Shimizu, K., Shijubo, N., … Watanabe, K. (2018). Distinct Phenotypes of Smokers with Fixed Airflow Limitation Identified by Cluster Analysis of Severe Asthma. *Annals of the American Thoracic Society*, *15*(1), 33–41.

Konstantellou, E., Papaioannou, A. I., Loukides, S., Patentalakis, G., Papaporfyriou, A., Hillas, G., … Kostikas, K. (2015). Persistent airflow obstruction in patients with asthma: Characteristics of a distinct clinical phenotype. *Respiratory Medicine*, *109*(11), 1404–1409.

Kostikas, K., Bakakos, P., Papiris, S., Stolz, D., & R. Celli, B. (2013). Systemic Biomarkers in the Evaluation and Management of COPD Patients: Are We Getting Closer to Clinical Application? *Current Drug Targets*, *14*(2), 177–191.

Kraft, M. (2011). Asthma Phenotypes and Interleukin-13 — Moving Closer to Personalized Medicine. *New England Journal of Medicine*, *365*(12), 1141–1144.

Kurukulaaratchy, R. J., Zhang, H., Patil, V., Raza, A., Karmaus, W., Ewart, S., & Arshad, S. H. (2015). Identifying the heterogeneity of young adult rhinitis through cluster analysis in the Isle of Wight birth cohort. *Journal of Allergy and Clinical Immunology*, *135*(1), 143–150.e7.

Lamprecht, B., McBurnie, M. A., Vollmer, W. M., Gudmundsson, G., Welte, T., Nizankowska-Mogilnicka, E., … Buist, S. A. (2011). COPD in never smokers: Results from the population-based burden of obstructive lung disease study. *Chest*, *139*(4), 752–763.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent Class Analysis With Distal Outcomes: A Flexible Model-Based Approach. *Structural Equation Modeling, 20*(1), 1–26.

Larsen, F. B., Pedersen, M. H., Friis, K., Glümer, C., & Lasgaard, M. (2017). A Latent Class Analysis of Multimorbidity and the Relationship to Socio-Demographic Factors and Health-Related Quality of Life. A National Population-Based Study of 162,283 Danish Adults. *PLoS One*, *12*(1), e0169426.

Lee, J. H., Rhee, C. K., Kim, K., Kim, J.-A., Kim, S. H., Yoo, K.H., … Jung, K.S. (2017). Identification of subtypes in subjects with mild-to-moderate airflow limitation and its clinical and socioeconomic implications. *International Journal of Chronic Obstructive Pulmonary Disease*, *12*, 1135–1144.

Lefaudeux, D., De Meulder, B., Loza, M. J., Peffer, N., Rowe, A., Baribaud, F., … Zwinderman, K. (2017). U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. *Journal of Allergy and Clinical Immunology*, *139*(6), 1797–1807.

Leiria, L. O. S., Martins, M. A., & Saad, M. J. A. (2015). Obesity and asthma: beyond TH2 inflammation. *Metabolism*, *64*(2), 172–181.

Lemanske, R. F., & Busse, W. W. (2003). 6. Asthma. *Journal of Allergy and Clinical Immunology*, *111*(2), S502–S519.

Lemiere, C., NGuyen, S., Sava, F., D'Alpaos, V., Huaux, F., & Vandenplas, O. (2014). Occupational asthma phenotypes identified by increased fractional exhaled nitric oxide after exposure to causal agents. *Journal of Allergy and Clinical Immunology*, *134*(5), 1063–1067.

Leventhal, A. M., Huh, J., & Dunton, G. F. (2014). Clustering of modifiable biobehavioral risk factors for chronic disease in US adults: a latent class analysis. *Perspectives in Public Health*, *134*(6), 331–338.

Leynaert, B., Neukirch, C., Kony, S., Guénégou, A., Bousquet, J., Aubier, M., & Neukirch, F. (2004). Association between asthma and rhinitis according to atopic sensitization in a population-based study. *Journal of Allergy and Clinical*

Immunology, *113*(1), 86–93.

Lim, H., Kwon, H. J., Lim, J. A., Choi, J. H., Ha, M., Hwang, S. S., & Choi, W. J. (2016). Short-term Effect of Fine Particulate Matter on Children's Hospital Admissions and Emergency Department Visits for Asthma: A Systematic Review and Meta-analysis. *Journal of Preventive Medicine & Public Health, 49*(4), 205-219.

Linnaeus, C. (1759). *Genera Morborum [Varieties of Disease]*. (Steinert, Ed.). Uppsala. Retrieved from http://linnean-online.org/120140.

Lloyd, C. M., & Saglani, S. (2013). T cells in asthma: influences of genetics, environment, and T-cell plasticity. *Journal of Allergy and Clinical Immunology, 131*(5), 1267-1274; quiz 1275.

Loureiro, C. C., Sa-Couto, P., Todo-Bom, A., & Bousquet, J. (2015). Cluster analysis in phenotyping a Portuguese population. *Revista Portuguesa de Pneumologia*, *21*(6), 299–306.

Loza, M. J., Adcock, I., Auffray, C., Chung, K. F., Djukanovic, R., Sterk, P. J., … Silkoff, P. E. (2016). Longitudinally Stable, Clinically Defined Clusters of Patients with Asthma Independently Identified in the ADEPT and U-BIOPRED Asthma Studies. *Annals of the American Thoracic Society*, *13 Suppl 1*, S102-3.

Loza, M. J., Djukanovic, R., Chung, K. F., Horowitz, D., Ma, K., Branigan, P., … Baribaud, F. (2016). Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. *Respiratory Research*, *17*(1), 165.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., … Murray, C. J. L. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, *380*(9859), 2095–2128.

Macfarlane, A. J., Kon, O. M., Smith, S. J., Zeibecoglou, K., Khan, L. N., Barata, L. T., … Kay, A. B. (2000). Basophils, eosinophils, and mast cells in atopic and nonatopic asthma and in late-phase allergic reactions in the lung and skin. *Journal of Allergy and Clinical Immunology*, *105*(1), 99–107.

Makikyro, E. M. S., Jaakkola, M. S., & Jaakkola, J. J. K. (2017). Subtypes of asthma based on asthma control and severity: a latent class analysis. *Respiratory Research*, *18*(1).

Malinovschi, A., Fonseca, J. A., Jacinto, T., Alving, K., & Janson, C. (2013). Exhaled nitric oxide levels and blood eosinophil counts independently associate with

wheeze and asthma events in National Health and Nutrition Examination Survey subjects. *Journal of Allergy and Clinical Immunology*, *132*(4), 821–827.e5.

Malinovschi, A., Janson, C., Borres, M., & Alving, K. (2016). Simultaneously increased fraction of exhaled nitric oxide levels and blood eosinophil counts relate to increased asthma morbidity. *Journal of Allergy and Clinical Immunology*, *138*(5), 1301-1308.e2.

Marciniuk, D., Ferkol, T., Nana, A., Oca, M. M. de, Rabe, K., Billo, N., & Zar, H. (2014). Respiratory diseases in the world. Realities of today – opportunities for tomorrow. *African Journal of Respiratory Medicine, 9*(1), 4-13.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is More Ever Too Much? The Number of Indicators per Factor in Confirmatory Factor Analysis. *Multivariate Behavioral Research*, *33*(2), 181–220.

Marshall, R. J. (2001). The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology*, *54*(6), 603–609.

Máspero, J. (2017). Reslizumab in the treatment of inadequately controlled asthma in adults and adolescents with elevated blood eosinophils: clinical trial evidence and future prospects. *Therapeutic Advances in Respiratory Disease*, *11*(8), 311–325.

McCreanor, J., Cullinan, P., Nieuwenhuijsen, M. J., Stewart-Evans, J., Malliarou, E., Jarup, L., ... Zhang, J. (2007). Respiratory effects of exposure to diesel traffic in persons with asthma. *The New England Journal of Medicine, 357*(23), 2348-2358.

McGrath, K. W., Icitovic, N., Boushey, H. A., Lazarus, S. C., Sutherland, E. R., Chinchilli, V. M., & Fahy, J. V. (2012). A Large Subgroup of Mild-to-Moderate Asthma Is Persistently Noneosinophilic. *American Journal of Respiratory and Critical Care Medicine*, *185*(6), 612–619.

Mendy, A., Forno, E., Niyonsenga, T., Carnahan, R., & Gasana, J. (2018). Prevalence and features of asthma-COPD overlap in the United States 2007-2012. *The Clinical Respiratory Journal*, *12*(8), 2369–2377.

Mesidor, M., Benedetti, A., El-Zein, M., Menzies, D., Parent, M. E., & Rousseau, M. C. (2019). Asthma phenotypes based on health services use for allergic diseases in a province-wide birth cohort. *Annals of Allergy Asthma & Immunology*, *122*(1), 50

Mihalcea, R., & Tarau, P. (2004) Textrank: Bringing order into texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* 404–411.

Miller, M. R. (2005). Standardisation of spirometry. *European Respiratory Journal*, *26*(2), 319–338.

Miralles, F., Gomez-Cabrero, D., Lluch-Ariet, M., Tegnér, J., Cascante, M., & Roca, J. (2014). Predictive medicine: outcomes, challenges and opportunities in the Synergy-COPD project. *Journal of Translational Medicine*, *12*(Suppl 2), S12.

Mølgaard, E., Thomsen, S. F., Lund, T., Pedersen, L., Nolte, H., & Backer, V. (2007). Differences between allergic and nonallergic rhinitis in a large sample of adolescents and adults. *Allergy*, *62*(9), 1033–1037.

Moore, W. C., Hastie, A. T., Li, X., Li, H., Busse, W. W., Jarjour, N. N., … Bleecker, E. R. (2014). Sputum neutrophil counts are associated with more severe asthma phenotypes using cluster analysis. *Journal of Allergy and Clinical Immunology*, *133*(6), 1557–1563.e5.

Moore, W. C., Meyers, D. A., Wenzel, S. E., Teague, W. G., Li, H., Li, X., … Bleecker, E. R. (2010). Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *American Journal of Respiratory and Critical Care Medicine*, *181*(4), 315–323.

Morais-Almeida, M., Pite, H., Pereira, A. M., Todo-Bom, A., Nunes, C., Bousquet, J., & Fonseca, J. (2013a). Prevalence and classification of rhinitis in the elderly: a nationwide survey in Portugal. *Allergy, 68*(9), 1150-1157.

Morais-Almeida, M., Santos, N., Pereira, A. M., Branco-Ferreira, M., Nunes, C., Bousquet, J., & Fonseca, J. A. (2013b). Prevalence and classification of rhinitis in preschool children in Portugal: a nationwide study. *Allergy, 68*(10), 1278-1288.

Muraro, A., Lemanske, R. F., Hellings, P. W., Akdis, C. A., Bieber, T., Casale, T. B., … Agache, I. (2016). Precision medicine in patients with allergic diseases: Airway diseases and atopic dermatitis—PRACTALL document of the European Academy of Allergy and Clinical Immunology and the American Academy of Allergy, Asthma & Immunology. *Journal of Allergy and Clinical Immunology*, *137*(5), 1347–1358.

Musk, A. W., Knuiman, M., Hunter, M., Hui, J., Palmer, L. J., Beilby, J., … James, A. (2011). Patterns of airway disease and the clinical diagnosis of asthma in the Busselton population. *European Respiratory Journal*, *38*(5), 1053–1059.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nathan, R. A., Dalal, A. A., Stanford, R. H., Meltzer, E. O., Schatz, M., Derebery, J., … DiBenedetti, D. B. (2010). Qualitative Development of the Rhinitis Control Assessment Test (RCAT), an Instrument for Evaluating Rhinitis Symptom Control. *The Patient: Patient-Centered Outcomes Research*, *3*(2), 91–99.

National Research Council. (2011). *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. *The National Academies Press*. Washington, D.C., USA: National Academies Press.

Newby, C., Heaney, L. G., Menzies-Gow, A., Niven, R. M., Mansur, A., Bucknall, C., … Brightling, C. (2014). Statistical cluster analysis of the british thoracic society severe refractory asthma registry: Clinical outcomes and phenotype stability. *PLoS One, 9*(7), e102987.

Ning, P., Guo, Y. F., Sun, T. Y., Zhang, H. S., Chai, D., & Li, X. M. (2016). [Study of the clinical phenotype of symptomatic chronic airways disease by hierarchical cluster analysis and two-step cluster analyses]. *Zhonghua nei ke za zhi*, *55*(9), 679–683.

Nunes, C., Pereira, A. M., & Morais-Almeida, M. (2017). Asthma costs and social impact. *Asthma Research and Practice*, *3*(1).

O'Byrne, P. M., Pedersen, S., Lamm, C. J., Tan, W. C., & Busse, W. W. (2009). Severe Exacerbations and Decline in Lung Function in Asthma. *American Journal of Respiratory and Critical Care Medicine*, *179*(1), 19–24.

Oberski D. (2016) Mixture Models: Latent Profile and Latent Class Analysis. In: Robertson J., Kaptein M. (Eds.) *Modern Statistical Methods for HCI*. Human–Computer Interaction Series. (pp. 275-287): Springer, Cham.

Oertzen, T., Hertzog, C., Lindenberger, U., & Ghisletta, P. (2010). The effect of multiple indicators on the power to detect inter-individual differences in change. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 627–646.

Onder, G., Palmer, K., Navickas, R., Jurevičienė, E., Mammarella, F., Strandzheva, M., … Marengoni, A. (2015). Time to face the challenge of multimorbidity. A European perspective from the joint action on chronic diseases and promoting healthy ageing across the life cycle (JA-CHRODIS). *European Journal of Internal Medicine*, *26*(3), 157–159.

Orie, N. G., Sluiter, H. J., De Vries, K., Tammeling, G.J., & Witkop, J. (1961). The host factor in bronchitis. In: Orie N.G. & Sluiter H.J. (Eds.), *Bronchitis* (pp. 43–59). Assen, The Netherlands: Royal Van Gorcum.

Papadopoulos, N. G., Bernstein, J. A., Demoly, P., Dykewicz, M., Fokkens, W., Hellings, P. W., … Cox, L. S. (2015). Phenotypes and endotypes of rhinitis and their impact on management: a PRACTALL report. *Allergy*, *70*(5), 474–494.

Park, H.W., Song, W.J., Kim, S.H., Park, H.K., Kim, S.H., Kwon, Y. E., … Cho, S.H. (2015). Classification and implementation of asthma phenotypes in elderly patients. *Annals of Allergy, Asthma & Immunology*, *114*(1), 18–22.

Park, S. Y., Baek, S., Kim, S., Yoon, S.Y., Kwon, H.S., Chang, Y.S., … Kim, T.B. (2013). Clinical significance of asthma clusters by longitudinal analysis in Korean asthma cohort. *PloS One*, *8*(12), e83540.

Patelis, A., Gunnbjörnsdottir, M., Malinovschi, A., Matsson, P., Önell, A., Högman, M., … Janson, C. (2012). Population-based study of multiplexed IgE sensitization in relation to asthma, exhaled nitric oxide, and bronchial responsiveness. *Journal of Allergy and Clinical Immunology*, *130*(2), 397–402.e2.

Pavord, I. D., Beasley, R., Agusti, A., Anderson, G. P., Bel, E., Brusselle, G., … Bush, A. (2018). After asthma: redefining airways diseases. *The Lancet*, *391*(10118), 350–400.

Pavord, I. D., Korn, S., Howarth, P., Bleecker, E. R., Buhl, R., Keene, O. N., … Chanez, P. (2012). Mepolizumab for severe eosinophilic asthma (DREAM): A multicentre, double-blind, placebo-controlled trial. *The Lancet*, *380*(9842), 651–659.

Pavord, I. D., Sterk, P. J., Hargreave, F. E., Kips, J. C., Inman, M. D., Louis, R., ... Djukanovic, R. (2002). Clinical applications of assessment of airway inflammation using induced sputum. *European Respiratory Journal*, 20 (Suppl 37), 40s-43s.

Pawankar, R., Canonica, G.W., Holgate, S.T., Lockey, R.F., Blaiss, M.S. World Allergy Organization (WAO) white book on allergy: update 2013. Milwaukee, WI: WAO; 2013.

Pellegrino, R. (2005). Interpretative strategies for lung function tests. *European Respiratory Journal*, *26*(5), 948–968.

Pembrey, L., Barreto, M. L., Douwes, J., Cooper, P., Henderson, J., Mpairwe, H., ... Pearce, N. (2018). Understanding asthma phenotypes: the World Asthma Phenotypes (WASP) international collaboration. *ERJ Open Research, 4*(3).

Pepe, M. S., & Janes, H. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics*, *8*(2), 474–484.

Prosperi, M. C. F., Sahiner, U. M., Belgrave, D., Sackesen, C., Buchan, I. E., Simpson, A., … Custovic, A. (2013). Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *American Journal of Respiratory and Critical Care Medicine*, *188*(11), 1303–1312.

Qiu, R., Xie, J., Chung, K. F., Li, N., Yang, Z., He, M., … Zhang, Q. (2018). Asthma Phenotypes Defined From Parameters Obtained During Recovery From a Hospital-Treated Exacerbation. *The Journal of Allergy and Clinical Immunology: In Practice*, *6*(6), 1960–1967.

Quanjer, P. H., Stanojevic, S., Cole, T. J., Baur, X., Hall, G. L., Culver, B. H., … Schindler, C. (2012). Multi-ethnic reference values for spirometry for the 3-95-yr age range: The global lung function 2012 equations. *European Respiratory Journal*, *40*(6), 1324–1343.

Rackemann, F. (1927). A clinical survey of 1074 patients with asthma followed for two years. *Journal of Laboratory and Clinical Medicine*, *12*, 1185–1197.

Reddel, H. K., Bateman, E. D., Becker, A., Boulet, L. P., Cruz, A. A., Drazen, J. M., … FitzGerald, J. M. (2015). A summary of the new GINA strategy: a roadmap to asthma control. *European Respiratory Journal*, *46*(3), 622–639.

Rice, J. P., Saccone, N. L., & Rasmussen, E. (2001). Definition of the phenotype. *Advances in Genetics, 42*, 69–76.

Robinson, D., Humbert, M., Buhl, R., Cruz, A. A., Inoue, H., Korom, S., … Nair, P. (2017). Revisiting Type 2-high and Type 2-low airway inflammation in asthma: current knowledge and therapeutic implications. *Clinical & Experimental Allergy*, *47*(2), 161–175.

Robinson, P. N. (2012). Deep phenotyping for precision medicine. *Human Mutation*, *33*(5), 777–780.

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

Rosado-Pinto, J., & Carreiro-Martins, P. (2017). The global alliance against chronic respiratory diseases Portugal: 10 years of experience. *Journal of Thoracic Disease*, *9*(10), 4132–4134.

Rosario, N., & Bielory, L. (2011). Epidemiology of allergic conjunctivitis. *Current Opinion in Allergy and Clinical Immunology*, *11*(5), 471–476.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010): Automatic Keyword Extraction from Individual Documents. In Berry M. W. & Kogan J. (Eds.), *Text Mining: Theory and Applications* (pp. 3-20). John Wiley & Sons, Inc.

Rossi, A., Butorac Petanjek, B., Chilosi, M., G Cosio, B., Flezar, M., Koulouris, N., … Miravitlles, M. (2017). Chronic obstructive pulmonary disease with mild airflow limitation: current knowledge and proposal for future research – a consensus document from six scientific societies. *International Journal of Chronic Obstructive Pulmonary Disease*, *Volume 12*, 2593–2610.

Rovner, G., Vowles, K. E., Gerdle, B., & Gillanders, D. (2015). Latent Class Analysis of the Short and Long Forms of the Chronic Pain Acceptance Questionnaire: Further Examination of Patient Subgroups. *Journal of Pain, 16*(11), 1095-1105.

Rufo, J., Taborda-Barata, L., & Lourenço, O. (2013). Serum biomarkers in elderly asthma. *Journal of Asthma*, *50*(10), 1011–1019.

Sá-Sousa, A., Amaral, R., Morais-Almeida, M., Araújo, L., Azevedo, L. F., Bugalho-Almeida, A., … Fonseca, J. A. (2015). Asthma control in the Portuguese National Asthma Survey. *Revista Portuguesa de Pneumologia (English Edition)*, *21*(4), 209–213.

Sá-Sousa, A., Fonseca, J. A., Pereira, A. M., Ferreira, A., Arrobas, A., Mendes, A., … Calvo, T. (2018). The Portuguese Severe Asthma Registry: Development, Features, and Data Sharing Policies. *BioMed Research International, 2018,* 1495039.

Sá-Sousa, A., Jacinto, T., Azevedo, L., Morais-Almeida, M., Robalo-Cordeiro, C., Bugalho-Almeida, A., … Fonseca, J. (2014). Operational definitions of asthma in recent epidemiological studies are inconsistent. *Clinical and Translational Allergy*, *4*(1), 24.

Sá-Sousa, A., Morais-Almeida, M., Azevedo, L., Carvalho, R., Jacinto, T., Todo-Bom, A., … Fonseca, J. (2012). Prevalence of asthma in Portugal - The Portuguese National Asthma Survey. *Clinical and Translational Allergy*, *2*(1), 15.

Sá-Sousa, A., Pereira, A. M., Almeida, R., Araújo, L., Couto, M., Jacinto, T., … Fonseca, J. A. (2019). Adult Asthma Scores—Development and Validation of Multivariable Scores to Identify Asthma in Surveys. *The Journal of Allergy and Clinical Immunology: In Practice*, *7*(1), 183–190.e6.

Sano, H., Iwanaga, T., Nishiyama, O., Sano, A., Higashimoto, Y., Tomita, K., & Tohda,

Y. (2016). Characteristics of phenotypes of elderly patients with asthma. *Allergology International*, *65*(2), 204–209.

Sauve, M., & Tuleau-Malot, C. (2014). Variable selection through CART. *ESAIM: Probability and Statistics*, *18*, 770–798.

Schatz, M., Hsu, J. W. Y., Zeiger, R. S., Chen, W., Dorenbaum, A., Chipps, B. E., & Haselkorn, T. (2014). Phenotypes determined by cluster analysis in severe or difficult-to-treat asthma. *Journal of Allergy and Clinical Immunology*, *133*(6), 1549–1556.

Schleich, F., Brusselle, G., Louis, R., Vandenplas, O., Michils, A., Pilette, C., … Joos, G. (2014). Heterogeneity of phenotypes in severe asthmatics. The Belgian Severe Asthma Registry (BSAR). *Respiratory Medicine*, *108*(12), 1723–1732.

Schleich, F., Corhay, J.-L., & Louis, R. (2016). Blood eosinophil count to predict bronchial eosinophilic inflammation in COPD. *European Respiratory Journal*, *47*(5), 1562–1564.

Schleich, F., Demarche, S., & Louis, R. (2016). Biomarkers in the Management of Difficult Asthma. *Current Topics in Medicinal Chemistry, 16*(14), 1561-1573.

Seino, Y., Hasegawa, T., Koya, T., Sakagami, T., Mashima, I., Shimizu, N., … Kikuchi, T. (2018). A Cluster Analysis of Bronchial Asthma Patients with Depressive Symptoms. *Internal Medicine*, *57*(14), 1967–1975.

Sekiya, K., Nakatani, E., Fukutomi, Y., Kaneda, H., Iikura, M., Yoshida, M., … Tanaka, H. (2016). Severe or life-threatening asthma exacerbation: patient heterogeneity identified by cluster analysis. *Clinical and Experimental Allergy : Journal of the British Society for Allergy and Clinical Immunology*, *46*(8), 1043–1055.

Sendín-Hernández, M. P., Ávila-Zarza, C., Sanz, C., García-Sánchez, A., Marcos-Vadillo, E., Muñoz-Bellido, F. J., … Dávila, I. (2018). Cluster Analysis Identifies 3 Phenotypes within Allergic Asthma. *The Journal of Allergy and Clinical Immunology: In Practice*, *6*(3), 955–961.e1.

Serrano-Pariente, J., Rodrigo, G., Fiz, J. A., Crespo, A., & Plaza, V. (2015). Identification and characterization of near-fatal asthma phenotypes by cluster analysis. *Allergy*, *70*(9), 1139–1147.

Shaaban, R., Zureik, M., Soussan, D., Neukirch, C., Heinrich, J., Sunyer, J., … Leynaert, B. (2008). Rhinitis and onset of asthma: a longitudinal population-based study. *The Lancet*, *372*(9643), 1049–1057.

Silkoff, P. E. (2005). ATS/ERS Recommendations for Standardized Procedures for the Online and Offline Measurement of Exhaled Lower Respiratory Nitric Oxide and Nasal Nitric Oxide, 2005. *American Journal of Respiratory and Critical Care Medicine*, *171*(8), 912–930.

Silkoff, P. E., Laviolette, M., Singh, D., FitzGerald, J. M., Kelsen, S., Backer, V., … Chanez, P. (2017). Identification of airway mucosal type 2 inflammation by using clinical biomarkers in asthmatic patients. *Journal of Allergy and Clinical Immunology*, *140*(3), 710–719.

Simpson, A., Hekking, P. P., Shaw, D., Riley, J., Sousa, A., Fleming, L., … Montuschi, P. (2017). Cluster analysis of treatable traits in the U-BIOPRED adult severe asthma cohort. *European Respiratory Journal*, *50*: OA314.

Singh, D., Kolsum, U., Brightling, C. E., Locantore, N., Agusti, A., & Tal-Singer, R. (2014). Eosinophilic inflammation in COPD: prevalence and clinical characteristics. *European Respiratory Journal*, *44*(6), 1697–1700.

Siroux, V., Ballardini, N., Soler, M., Lupinek, C., Boudier, A., Pin, I., … Bousquet, J. (2018). The asthma-rhinitis multimorbidity is associated with IgE polysensitization in adolescents and adults. *Allergy*, *73*(7), 1447–1458.

Siroux, V., Basagana, X., Boudier, A., Pin, I., Garcia-Aymerich, J., Vesin, A., … Sunyer, J. (2011). Identifying adult asthma phenotypes using a clustering approach. *European Respiratory Journal*, *38*(2), 310–317.

Siroux, V., Boudier, A., Nadif, R., Lupinek, C., Valenta, R., & Bousquet, J. (2019). Association between asthma, rhinitis, and conjunctivitis multimorbidities with molecular IgE sensitization in adults. *Allergy, 74*(4), 824-827.

Soriano, J. B., Ancochea, J., Miravitlles, M., García-Río, F., Duran-Tauleria, E., Muñoz, L., … Sobradillo-Peña, V. (2010). Recent trends in COPD prevalence in Spain: A repeated cross-sectional survey 1997-2007. *European Respiratory Journal*, *36*(4), 758–765.

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunananthan, S., … Bergman, H. (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of Clinical Epidemiolog, 63*(6):638-46.

Stanojevic, S., Wade, A., Stocks, J., Hankinson, J., Coates, A. L., Pan, H., … Cole, T. J. (2008). Reference ranges for spirometry across all ages: A new approach.

American Journal of Respiratory and Critical Care Medicine, *177*(3), 253–260.

Sutherland, E. R., Goleva, E., King, T. S., Lehman, E., Stevens, A. D., Jackson, L. P., … Fahy, J. V. (2012). Cluster Analysis of Obesity and Asthma Phenotypes. *PLoS One, 7*(5), e36631

Tanaka, H., Nakatani, E., Fukutomi, Y., Sekiya, K., Kaneda, H., Iikura, M., … Taniguchi, M. (2018). Identification of patterns of factors preceding severe or life-threatening asthma exacerbations in a nationwide study. *Allergy*, *73*(5), 1110–1118.

The Lancet. (2006) [No authors listed]. A plea to abandon asthma as a disease concept. *The Lancet*, *368*(9537), 705.

Tilert, T., Dillon, C., Paulose-Ram, R., Hnizdo, E., & Doney, B. (2013). Estimating the U.S. prevalence of chronic obstructive pulmonary disease using pre- and post-bronchodilator spirometry: the National Health and Nutrition Examination Survey (NHANES) 2007–2010. *Respiratory Research*, *14*(1), 103.

Todo-Bom, A., Loureiro, C., Almeida, M. M., Nunes, C., Delgado, L., Castel-Branco, G., & Bousquet, J. (2007). Epidemiology of rhinitis in Portugal: evaluation of the intermittent and the persistent types. *Allergy*, *62*(9), 1038–1043.

Toelle, B. G., Xuan, W., Bird, T. E., Abramson, M. J., Atkinson, D. N., Burton, D. L., … Marks, G. B. (2013). Respiratory symptoms and illness in older Australians: the Burden of Obstructive Lung Disease (BOLD) study. *The Medical Journal of Australia*, *198*(3), 144–148.

Toren, K., Brisman, J., & Jarvholm, B. (1993). Asthma and asthma-like symptoms in adults assessed by questionnaires: A literature review. *Chest, 104*(2), 600-608.

Tran, T. N., Zeiger, R. S., Peters, S. P., Colice, G., Newbold, P., Goldman, M., & Chipps, B. E. (2016). Overlap of atopic, eosinophilic, and TH2-high asthma phenotypes in a general population with current asthma. *Annals of Allergy, Asthma & Immunology*, *116*(1), 37–42.

Turing, A. M. (1937). On Computable Numbers, With Application to the Entscheidungs Problem. *Proceedings of the London Mathematical Society, s2-42*(1), 230–265

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., & Roland, M. (2009). Defining comorbidity: Implications for understanding health and health services. *Annals of Family Medicine, 7*(4):357-63.

Van der Molen, T., Fletcher, M., & Price, D. (2018). Identifying Patient Attitudinal Clusters Associated with Asthma Control: The European REALISE Survey. *The Journal of Allergy and Clinical Immunology: In Practice*, *6*(3), 962–971.

Van Gerven, L., Steelant, B., & Hellings, P. W. (2018). Nasal hyperreactivity in rhinitis: A diagnostic and therapeutic challenge. *Allergy*, *73*(9), 1784–1791.

Vanfleteren, L. E. G. W., Kocks, J. W. H., Stone, I. S., Breyer-Kohansal, R., Greulich, T., Lacedonia, D., … Agusti, A. (2014). Moving from the Oslerian paradigm to the post-genomic era: are asthma and COPD outdated terms? *Thorax*, *69*(1), 72–79.

Veldhoen, M., Uyttenhove, C., van Snick, J., Helmby, H., Westendorf, A., Buer, J., ... Stockinger, B. (2008). Transforming growth factor-beta 'reprograms' the differentiation of T helper 2 cells and promotes an interleukin 9-producing subset. *Nature Immunology, 9*(12), 1341-1346.

Vogelmeier, C. F., Criner, G. J., Martinez, F. J., Anzueto, A., Barnes, P. J., Bourbeau, J., … Agustí, A. (2017). Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *American Journal of Respiratory and Critical Care Medicine*, *195*(5), 557–582.

Vollmer, W. M., Gíslason, P., Burney, P., Enrigh, P. L., Gulsvik, A., Kocabas, A., & Buist, A. S. (2009). Comparison of spirometry criteria for the diagnosis of COPD: Results from the BOLD study. *European Respiratory Journal*, *34*(3), 588–597.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., & for the STROBE initiative. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guildelines for reporting observational studies. *Journal of Clinical Epidemiology*, *61*(4), 344–349.

Wadsworth, S., Sin, D., & Dorscheid. (2011). Clinical update on the use of biomarkers of airway inflammation in the management of asthma. *Journal of Asthma and Allergy*, 77.

Wagener, A. H., de Nijs, S. B., Lutter, R., Sousa, A. R., Weersink, E. J. M., Bel, E. H., & Sterk, P. J. (2015). External validation of blood eosinophils, FE NO and serum periostin as surrogates for sputum eosinophils in asthma. *Thorax*, *70*(2), 115–120.

Wagner, J. A. (2002). Overview of Biomarkers and Surrogate Endpoints in Drug Development. *Disease Markers*, *18*(2), 41–46.

Wang, J., & Wang, X. (2012). *Structural Equation Modeling: Applications Using Mplus*.

Chichester: Wiley.

Wang, L., Liang, R., Zhou, T., Zheng, J., Liang, B. M., Zhang, H. P., … Wang, G. (2017). Identification and validation of asthma phenotypes in Chinese population using cluster analysis. *Annals of Allergy, Asthma & Immunology*, *119*(4), 324–332.

Wardlaw, A. J., Silverman, M., Siva, R., Pavord, I. D., & Green, R. (2005). Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clinical and Experimental Allergy : Journal of the British Society for Allergy and Clinical Immunology*, *35*(10), 1254–1262.

Weatherall, M., Travers, J., Shirtcliffe, P. M., Marsh, S. E., Williams, M. V, Nowitz, M. R., … Beasley, R. (2009). Distinct clinical phenotypes of airways disease defined by cluster analysis. *European Respiratory Journal*, *34*(4), 812–818

Wenzel, S. (2006). Asthma: defining of the persistent adult phenotypes. *The Lancet*, *368*(9537), 804-813.

Wenzel, S., Castro, M., Corren, J., Maspero, J., Wang, L., Zhang, B., ... Teper, A. (2016). Dupilumab efficacy and safety in adults with uncontrolled persistent asthma despite use of medium-to-high-dose inhaled corticosteroids plus a long-acting beta2 agonist: a randomised double-blind placebo-controlled pivotal phase 2b dose-ranging trial. *The Lancet*, 388(10039), 31-44.

Wenzel, S. E. (2012). Asthma phenotypes: the evolution from clinical to molecular approaches. *Nature Medicine*, *18*(5), 716–725.

Wenzel, S., Ford, L., Pearlman, D., Spector, S., Sher, L., Skobieranda, F., … Pirozzi, G. (2013). Dupilumab in Persistent Asthma with Elevated Eosinophil Levels. *New England Journal of Medicine*, *368*(26), 2455–2466.

WHO. (1995). Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. *World Health Organization Technical Report Series*. Retrieved from: https://apps.who.int/iris/bitstream/handle/10665/37003/WHO_TRS_854.pdf;jsessionid=4995DA17EA305C785A2F66BB7C69225A?sequence=1

WHO. (2007). Global surveillance, prevention and control chronic respiratory diseases. *World Health Organization*. Retrieved from: https://www.who.int/gard/publications/GARD%20Book%202007.pdf

WHO. (2008). 2008-2013 Action Plan for the Global Strategy for the Prevention and Control of Noncommunicable Diseases: prevent and control cardiovascular

diseases, cancers, chronic respiratory diseases and diabetes. *World Health Organization*. Retrieved from: https://www.who.int/nmh/publications/ncd_action_plan_en.pdf

Wills-Karp, M. (1999). Immunologic basis of antigen-induced airway hyperresponsiveness. *Annual Review of Immunology*, *17*(1), 255–281.

Wise, S. K., Lin, S. Y., Toskala, E., Orlandi, R. R., Akdis, C. A., Alt, J. A., … Zacharek, M. (2018). International Consensus Statement on Allergy and Rhinology: Allergic Rhinitis. *International Forum of Allergy & Rhinology*, *8*(2), 108–352.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.): Elsevier.

Woodruff, P. G., Modrek, B., Choy, D. F., Jia, G., Abbas, A. R., Ellwanger, A., … Fahy, J. V. (2009). T-helper type 2-driven inflammation defines major subphenotypes of asthma. *American Journal of Respiratory and Critical Care Medicine*, *180*(5), 388–395.

Weatherall, M., Travers, J., Shirtcliffe, P. M., Marsh, S. E., Williams, M. V, Nowitz, M. R., … Beasley, R. (2009). Distinct clinical phenotypes of airways disease defined by cluster analysis. *European Respiratory Journal*, *34*(4), 812–818.

Wu, D., Bleier, B. S., Li, L., Zhan, X., Zhang, L., Lv, Q., … Wei, Y. (2018). Clinical Phenotypes of Nasal Polyps and Comorbid Asthma Based on Cluster Analysis of Disease History. *The Journal of Allergy and Clinical Immunology: In Practice*, *6*(4), 1297–1305.e1.

Wu, W., Bleecker, E., Moore, W., Busse, W. W., Castro, M., Chung, K. F., … Wenzel, S. E. (2014). Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. *Journal of Allergy and Clinical Immunology*, *133*(5), 1280–1288.

Ye, W.J., Xu, W.G., Guo, X.J., Han, F.F., Peng, J., Li, X.M., … Liu, S. (2017). Differences in airway remodeling and airway inflammation among moderate-severe asthma clinical phenotypes. *Journal of Thoracic Disease*, *9*(9), 2904–2914.

Youroukova, V. M., Dimitrova, D. G., Valerieva, A. D., Lesichkova, S. S., Velikova, T. V, Ivanova-Todorova, E. I., & Tumangelova-Yuzeir, K. D. (2017). Phenotypes Determined by Cluster Analysis in Moderate to Severe Bronchial Asthma. *Folia Medica*, *59*(2), 165–173.

Zaihra, T., Walsh, C. J., Ahmed, S., Fugère, C., Hamid, Q. A., Olivenstein, R., …

Benedetti, A. (2016). Phenotyping of difficult asthma using longitudinal physiological and biomarker measurements reveals significant differences in stability between clusters. *BMC Pulmonary Medicine*, *16*(1), 74.

Zedan, M. M. (2015). Clinical asthma phenotyping: A trial for bridging gaps in asthma management. *World Journal of Clinical Pediatrics*, *4*(2), 13.

7. Appendices

Appendix I

Characterization of asthma phenotypes derived from data-driven methods: Systematic review and unsupervised exploratory analysis *(draft)*

R. Amaral, T. Jacinto, B. Sousa-Pinto, J.A. Fonseca

Aims

To review asthma phenotypes derived from data-driven methods using variables easily measurable in a clinical setting, and to characterize them with unsupervised exploratory methods (automatic keyword extraction, multiple correspondence analysis and hierarchical clustering).

Methods

Three databases were used to perform the systematic search: PubMed, Scopus, and Web of Science, with no date/language restrictions. Studies were considered eligible if they identify asthma subtypes/phenotypes of adult patients (≥ 18 years old), using clinical parameters/variables that can be collected in current clinical practice (such as demographic, clinical, inflammatory, functional, healthcare use, use of asthma-medication) and that applied data-driven methods. Studies were excluded if they: a) focused exclusively on children and/or genotyping and theory-driven methods and were b) conference abstracts, editorials and opinion articles with no original data. Reviews were initially included to explore the references. Non-English publications were translated if considered eligible. Finally, only full-text studies containing original data were included. The detailed search strategy is available in the Table 19. Studies were assessed by two independent reviewers (R.A. & T.J.) independently and disagreements were resolved by consensus or by a third reviewer (J.A.F.). Unweighted kappa statistics was calculated.

Table 19. Search strategy to retrieve studies according to each database

Database	Query
PubMed	(phenotyp*[Title/Abstract] OR cluster*[Title/Abstract]) AND ("Asthma"[MeSH] OR asthm*[Title/Abstract]) AND ("Adult"[MeSH] OR "Adult"[Title/Abstract] OR adult*[Title/Abstract] OR "Middle Aged"[Mesh:NoExp] OR "Aged"[Mesh:NoExp]) AND (humans[mesh:noexp]) OR (#3 NOT animals[mesh:noexp]) NOT ((Review[ptyp] OR Meta-Analysis[ptyp] OR Letter[ptyp] OR Case Reports[ptyp]))
Scopus	(TITLE-ABS-KEY (asthm*) AND TITLE-ABS-KEY ((phenotyp* OR cluster*)) AND TITLE-ABS-KEY ((adult* OR "middle aged" OR elderly))) AND (EXCLUDE (DOCTYPE, "re") OR EXCLUDE (DOCTYPE, "le") OR EXCLUDE (DOCTYPE, "ed") OR EXCLUDE (DOCTYPE, "no") OR EXCLUDE (DOCTYPE, "ch") OR EXCLUDE (DOCTYPE, "sh"))
Web of Science	TÓPICO: (asthm*) AND TÓPICO: ((phenotyp* OR cluster*)) AND TÓPICO: ((adult* or middle aged or elderly)) Refinado por: [excluindo] TIPOS DE DOCUMENTO: (BOOK CHAPTER OR REVIEW OR EDITORIAL MATERIAL OR NOTE OR LETTER)

Figure 15 summarizes the selection criteria for the included studies. We identified 6,415 citations, of which 3,757 unique published articles were identified after exclusion of duplicate articles. After screening titles and abstracts, 419 studies were found to be potentially eligible for further review. Fifty-two studies of data-driven asthma phenotype were included (Figure 15).

For data preprocessing, we extract the English text and annotate it for language detection. Annotation performs natural language processing (tokenisation, parts of speech tagging, lemmatisation and dependency parsing), to identify words and label if the word is a noun/verb/adverb or if it is a person, a city, a number or a specific object (Berry & Kogan, 2010).

For automatic keyword extraction we used both algorithms: a graph-based ranking algorithm – TextRank, that finds the most relevant keywords and their cooccurrence in a sentence (Mihalcea & Tarau, 2004); and an unsupervised, domain-independent, and language-independent method – rapid automatic keywords extraction (RAKE), that search for keywords by looking to a contiguous sequence of words which do not

contain irrelevant words, namely by calculating a score for each word which is part of any candidate keyword (Rose, Engel, Cramer & Cowley, 2010).

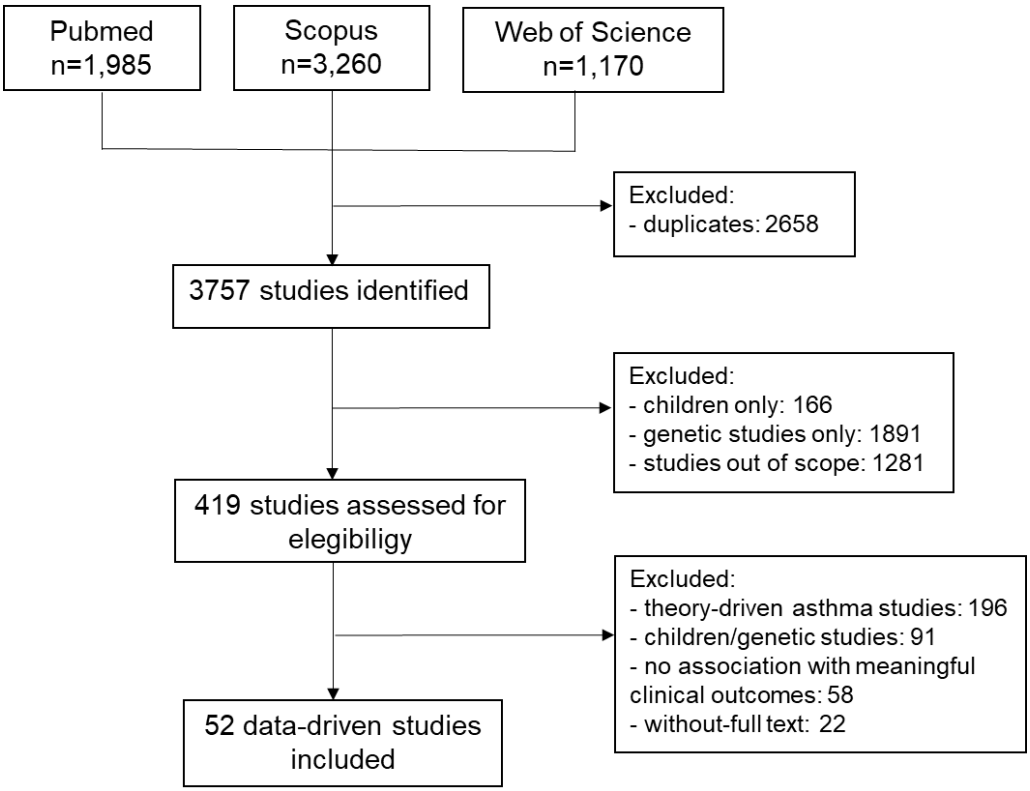


Figure 15. Flow diagram of articles selection process.

An exploratory analysis of the potential relationships between the multiple domains of asthma phenotypes, considering the different patient's characteristics, derived from difference data sources was then performed using a multiple correspondence analysis (MCA). MCA allows the analysis of multivariate categorical data and visualization of the results in a graphical manner (Greenacre, 1992; Sourial et al., 2010). For each study included in the systematic review, the socio-demographic, clinical, functional, inflammation, atopy, and other parameters (such as psychological/behavior variables, asthma-related medication use and healthcare use) were marked as a '1' if present and '0' if absent. The matrix data were then converted into dimensions that were structured from the most explicative to the least. To allow visualization, the scores from the two dimensions that account for the most variance are projected to create a factor plane.

Finally, a hierarchical clustering using the Euclidian distance and Ward's clustering method was performed, in order to identify the major groups of variables/characteristics retrieved from the asthma phenotypes, in an unsupervised manner.

All analyses were performed using the R programming environment (version 3.2.0) using the following packages: “udpipe”, “textrank”, “FactoMineR”, “cluster”, “factoextra” and “wordcloud”.

Preliminary Results

Keyword extraction

In Figure 16 is demonstrated the more frequent cooccurrence of keywords in the same sentence, in this case only nouns or adjectives, obtained with the TextRank algorithm. The most frequent are: asthma + onset (n=32), lung + function (n=23), severe + asthma (n=20), late + onset (n=19) and airflow + obstruction (n=15).

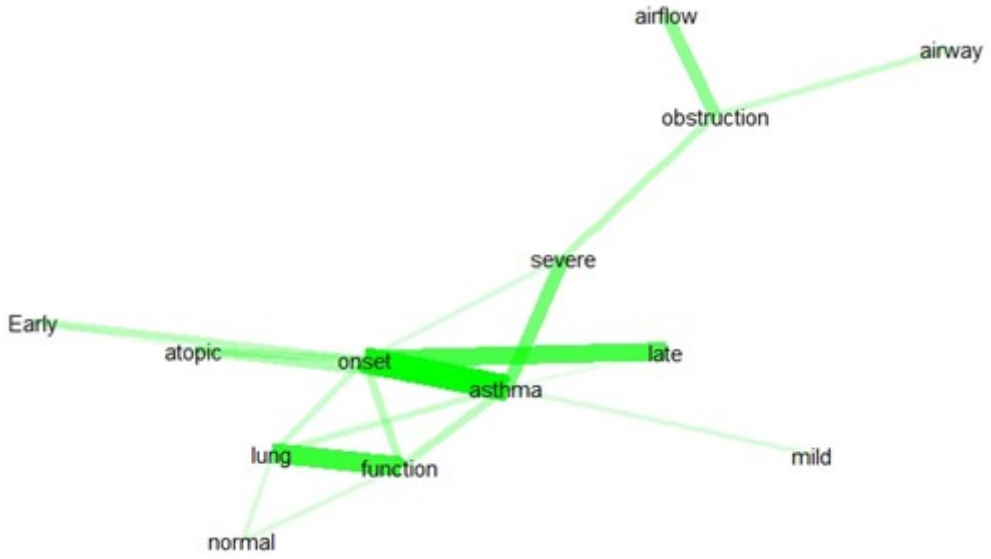


Figure 16. Keyword cooccurrence graph. The green line connects the cooccurrence of the keywords and the thickness represents the frequency of the cooccurrence.

Table 20 lists the top 15 keywords extracted by RAKE, order by the highest RAKE score to the lowest. The keywords that obtained a higher RAKE score were: “significant adherence barriers”, “minimal airflow obstruction”, “persistent airflow obstruction” (6.0, 5.7 and 5.7, respectively). However, when considering all obtained keywords and ordering by its frequency, combined keywords such as “normal lung function” (n=9), “onset atopic asthma” (n=8), “lung function” (n=8), “Severe asthma” (n=8) and “onset asthma” (n=8) were the most frequent (Figure 17).

Table 20. Top 15 of keywords extracted by RAKE algorithm.

Keywords	Freq	RAKE score
Significant adherence barriers	2	6
Minimal airflow obstruction	2	5.7
Persistent airflow obstruction	2	5.7
Lower lung function	2	5.5
Good lung function	2	5.2
Slight airway obstruction	2	5.2
Severe airflow obstruction	5	5
Normal lung function	9	4.75
Late disease onset	2	4.52
Low exacerbation risk	2	4.2
Onset allergic asthma	3	4.1
High blood eosinophils	2	4.2
Onset non-atopic asthma	2	3.8
Onset atopic asthma	8	3.6
COPD overlap group	2	3.5

Legend: RAKE: Rapid Automatic Keyword Extraction; COPD: chronic obstructive pulmonary disease; Freq: frequency

Figure 17. Word cloud of the more frequent combination of keywords obtained with RAKE algorithm.

Appendix II

Study I

Amaral, R., Fonseca, J. A., Jacinto, T., Pereira, A. M., Malinovschi, A., Janson, C., & Alving, K. (2018). Having concomitant asthma phenotypes is common and independently relates to poor lung function in NHANES 2007–2012. *Clinical and Translational Allergy*, *8*:13. doi:10.1186/s13601-018-0201-3.

Amaral et al. Clin Transl Allergy (2018) 8:13
https://doi.org/10.1186/s13601-018-0201-3

Clinical and Translational Allergy

RESEARCH Open Access

Having concomitant asthma phenotypes is common and independently relates to poor lung function in NHANES 2007–2012

Rita Amaral[1,2*], João A. Fonseca[1,3,4], Tiago Jacinto[1,2,4], Ana M. Pereira[1,4], Andrei Malinovschi[5], Christer Janson[6] and Kjell Alving[7]

Abstract

Background: Evidence for distinct asthma phenotypes and their overlap is becoming increasingly relevant to identify personalized and targeted therapeutic strategies. In this study, we aimed to describe the overlap of five commonly reported asthma phenotypes in US adults with current asthma and assess its association with asthma outcomes.

Methods: Data from the National Health and Nutrition Examination Surveys (NHANES) 2007–2012 were used (n = 30,442). Adults with current asthma were selected. Asthma phenotypes were: B-Eos-high [if blood eosinophils (B-Eos) ≥ 300/mm^3]; FeNO-high (FeNO ≥ 35 ppb); B-Eos&FeNO-low (B-Eos < 150/mm^3 and FeNO < 20 ppb); asthma with obesity (AwObesity) (BMI ≥ 30 kg/m^2); and asthma with concurrent COPD. Data were weighted for the US population and analyses were stratified by age (< 40 and ≥ 40 years old).

Results: Of the 18,619 adults included, 1059 (5.6% [95% CI 5.1–5.9]) had current asthma. A substantial overlap was observed both in subjects aged < 40 years (44%) and ≥ 40 years (54%). The more prevalent specific overlaps in both age groups were AwObesity associated with either B-Eos-high (15 and 12%, respectively) or B-Eos&FeNO-low asthma (13 and 11%, respectively). About 14% of the current asthma patients were "non-classified". Regardless of phenotype classification, having concomitant phenotypes was significantly associated with (adjusted OR, 95% CI) ≥ 2 controller medications (2.03, 1.16–3.57), and FEV_1 < LLN (3.21, 1.74–5.94), adjusted for confounding variables.

Conclusions: A prevalent overlap of commonly reported asthma phenotypes was observed among asthma patients from the general population, with implications for objective asthma outcomes. A broader approach may be required to better characterize asthma patients and prevent poor asthma outcomes.

Keywords: Asthma, Asthma-related outcomes, Epidemiological study, Overlap, Phenotypes

Background

Profiling asthma phenotypes is becoming increasingly relevant to choose the most appropriate therapeutic strategy for individual patients, and to provide optimal improvement of disease control and quality of life [1, 2].

The predominant pathophysiological mechanism of asthma is type 2-mediated, associated with atopy and eosinophilic inflammation [3, 4]. However, it has been shown that asthma is a heterogeneous disease that involves other mechanisms that are not so well understood and respond poorly to corticosteroid therapy (e.g. non-type 2-mediated) [3, 5, 6].

There has been a recent rise in the number of studies that try to identify asthma phenotypes based on non-invasive type 2-markers, such as blood eosinophils (B-Eos) count, fraction of exhaled nitric oxide (FeNO), serum IgE, and/or serum periostin [7–10]. Moreover, there appears to be an additive role of biomarkers, such as B-Eos and FeNO, in relation to recent asthma morbidity [7, 11, 12].

*Correspondence: rita.samaral@gmail.com
[1] CINTESIS- Center for Health Technology and Services Research, Faculty of Medicina, University of Porto, Edifício Nascente, Piso 2, Rua Dr. Plácido da Costa, s/n, 4200-450 Porto, Portugal
Full list of author information is available at the end of the article

However, there is little information regarding the appropriate use of these biomarkers in asthma phenotype classification, particularly when a significant overlap occurs. Also, the importance of having concomitant asthma phenotypes for disease outcomes has scarcely been studied in the general population. This information may be useful to identify personalized and targeted therapeutic strategies [13, 14].

Recently, an extensive overlap of asthma phenotypes was described [15]. However, only type 2-high, atopic, and eosinophilic asthma were examined. The extent of overlap with other phenotypes commonly reported in the literature, among adults with asthma from the general population remains unknown. Asthma phenotypes are frequently reported in the literature according to the high levels of systemic and local type 2-markers (B-Eos high and FeNO-high, respectively) [1–4, 16–18]. However, other distinct subgroups of asthma phenotypes are increasingly being reported due to its characteristics of steroid therapy resistance and lack of inflammatory markers: e.g. subjects with asthma without evidence of type 2 inflammation (Th2-low phenotype); obese asthmatic subjects (obesity-related asthma phenotype); and patients with asthma-COPD overlap syndrome [19–22]. Therefore, we hypothesized that if, in general population, occurs a high proportion of overlap of commonly reported asthma phenotypes, there may be a need for improving the definition of asthma phenotypes. Additionally, asthma subjects with multiple phenotypes may have poorer asthma-related outcomes.

The aims of this study were to describe the proportion of overlap of five commonly reported asthma phenotypes: asthma with obesity (AwObesity), asthma with concurrent COPD (AwCOPD), B-Eos-high, FeNO-high and B-Eos&FeNO-low asthma, and to examine the association of their overlap with asthma-related outcomes, using population-based data from the National Health and Nutrition Examination Surveys (NHANES), 2007–2012.

Methods

Study design

The NHANES is a nationally representative survey of the civilian, non-institutionalized U.S. population that uses a complex stratified, multistage probability sampling. Further details on NHANES survey design databases can be found in Additional file 1: Supplementary methods. The National Center for Health Statistics, Ethics Review Board approved NHANES protocol, and all participants gave written informed consent.

Subjects selection

Six survey years (NHANES 2007–2012) were analyzed, resulting in 30,442 individuals of all ages (Fig. 1). We included adults (≥ 18 years-old) with current asthma (n = 1059), defined by a positive answer to the questions: "Has a doctor ever told you that you have asthma?" together with "Do you still have asthma?", and either "wheezing/whistling in the chest in the past 12 months" or "asthma attack in the past 12 months."

Variables

Demographic characteristics, such as age, gender, body mass index (BMI), race/ethnicity, and educational status were analyzed. B-Eos count, FeNO and spirometric measurements, collected at the NHANES Mobile Examination Center were also examined. A detailed description of the procedures can be found elsewhere [23–25]. FeNO and spirometric measurements not fulfilling ATS/ERS recommendations [26, 27] were excluded (n = 653). After predicted values of basal FEV_1 and FEV_1/FVC were calculated [28], with a correction factor for ethnicity [29], abnormal lung function was defined if either one of them were less than the lower limit of normal (LLN), defined as lower fifth percentile of the reference population [30].

Prescription medications used last month were also analyzed [31]. More details regarding the inclusion of reliever and controller medications for asthma and the definitions of each asthma-related variable included in the analysis (asthma attack, asthma-related emergency department (ED) visit, work/school absenteeism, asthma symptoms, smoking status and rhinitis) are provided in the supplementary material (see Additional file 1: Supplementary methods).

Asthma phenotypes definition

A B-Eos count ≥ 300/mm³ was used to define an B-Eos-high asthma phenotype [32, 33], while FeNO-high was defined as FeNO ≥ 35 ppb [34]. Asthma patients with both B-Eos < 150/mm³ and FeNO < 20 ppb were categorized as B-Eos&FeNO-low asthma [35]. Additionally, we considered subjects with either B-Eos-high or FeNO-high as having "Type 2-high" asthma.

The AwObesity phenotype was defined by a BMI ≥ 30 kg/m² in individuals with current asthma [36]. Finally, the AwCOPD phenotype was considered if participants ≥ 40 years-old had concurrent asthma and COPD, defined by a positive answer to "Has a doctor ever told you that you have chronic bronchitis/emphysema", with age of diagnosis ≥ 40 years and having self-reported smoking history (being either a current or ex-smoker) [37, 38].

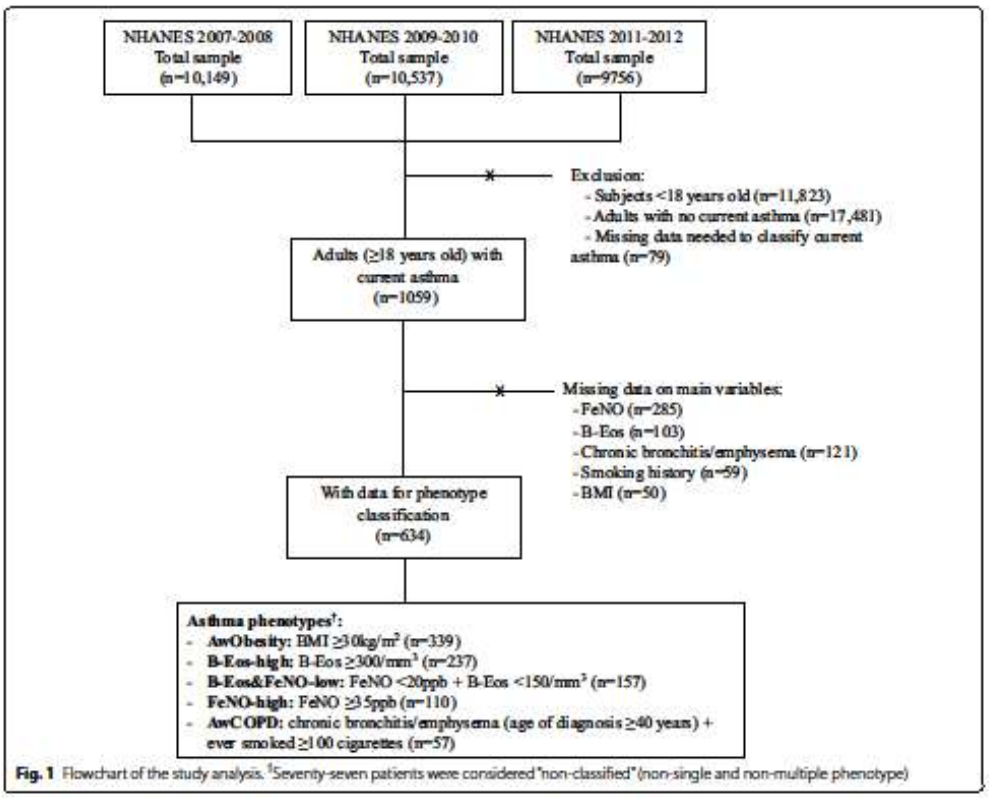


Fig. 1 Flowchart of the study analysis. †Seventy-seven patients were considered "non-classified" (non-single and non-multiple phenotype)

Statistical analysis

In accordance with the NHANES sampling design, the weights for each full sample 2-year mobile examination center were used to obtain weighted percentages adjusted to the US adult population.

Categorical variables were described as frequencies and weighted proportions, and continuous variables were described as median and first and third quartiles (Q1–Q3). Chi square test and Mann–Whitney U-test were used to compare groups. To explore the association of concomitant (having at least 2 concurrent) phenotypes with each asthma-related outcome we performed multivariate logistic regression modelling. Separate models were run using each asthma-related outcome and abnormal lung function as dependent variable and having multiple phenotypes as independent variables. Adjustments were also made for potential confounders: sex, age, race, current smoking and rhinitis. Adjusted odds ratios (aOR) with 95% confidence intervals (95% CI) were presented, and model fit was assessed using the *svylogitgof* function [39].

According to age (<40 or ≥40 years-old), a four- or five-set Venn-Euler diagram was used to quantify the proportion of individuals with different asthma phenotypes and to illustrate the overlap.

The diagrams were created using R software version 3.2.0 ("VennDiagram", "venneuler" and "reshape2" packages) and all statistical analyses were performed in Stata version 13.1 (StataCorp, TX, USA), using the *survey* command to account for the complex sampling design and weights in the NHANES. The *MI* command was used to perform sensitivity analysis by multiple-imputation of missing values; however, to create the Venn-Euler diagrams, a listwise deletion for missing data was applied. A p value < 0.05 was considered statistically significant.

Results

Of the 18,619 adults included in NHANES 2007–2012 datasets, 1059 (5.6% [95% CI 5.1–5.9]) had current asthma (Fig. 1). Of these, 63% were female, and the median (Q1–Q3) age was 48.0 (32.0–62.0) years. After excluding subjects with missing data on the main variables, 634 individuals were included for phenotype classification (Fig. 1). Despite having all information available, 77 patients did not meet the criteria for any of the defined asthma phenotypes and were considered "non-classified". These were non-obese subjects with asthma who did not meet the criteria for COPD, had B-Eos values ranging between 150 and 300/mm^3, and FeNO ranging 20–34 ppb.

Demographic characteristics of adults with current asthma included and excluded from the analysis and patients with single (n = 271) and multiple phenotypes (n = 286) are described in Table 1.

There is a female predominance in both groups (64 and 66%, respectively). Subjects with multiple phenotypes were older ($p = 0.003$), had higher BMI ($p < 0.001$), were more often obese ($p < 0.001$) and ex-smokers ($p = 0.003$), and a higher proportion of patients were treated with inhaled corticosteroids (ICS) ($p = 0.01$), than those with only one phenotype. Females were more obese, regardless the number of concomitant asthma phenotypes (data not shown).

Phenotypes and overlap description

The weighted proportions of asthma phenotypes were (in descending order): 49% for AwObesity, 36% for B-Eos-high asthma, 26% for B-Eos&FeNO-low asthma, 18% for FeNO-high asthma, and 8% for AwCOPD (Table 1).

Demographic and clinical characteristics among all 5 asthma phenotypes and the "non-classified" group are described in Table 2.

There is a female predominance among all phenotypes, particularly in the B-Eos&FeNO-low (78%). Subjects with AwCOPD phenotype were the oldest group (median [Q1–Q3]: 61.0 [52.0–69.0] years-old), with the lowest proportion of individuals that had ≥ high school and lowest FEV_1/FVC (0.63 [0.50–0.75]), comparing to the other phenotypes.

When categorized by age, < 40 (n = 227) and ≥ 40 years-old (n = 330), the most prevalent phenotypes were AwObesity (42 and 53%, respectively) and B-Eos-high asthma (34 and 37%). The less ones were FeNO-high asthma (18 and 19%) and AwCOPD (19% in the older group) (Fig. 2).

The areas of intersection in the four- and five-set Venn-Euler diagrams revealed 5 and 12 overlapping categories, and proportions of 17 and 12% of non-classified asthma subjects, respectively.

In both diagrams, a substantial total overlap was observed: 44% in subjects < 40 years-old and 54% in subjects ≥ 40 years-old. About 40% of the individuals in both age groups had two concomitant asthma phenotypes, 4% of the younger group had 3 concomitant phenotypes and 13% of the older group had ≥ 3 (Table 3 and Fig. 2). Furthermore, 1% of the older subjects had four concomitant asthma phenotypes: AwObesity, AwCOPD, FeNO-high, and B-Eos-high asthma.

The most prevalent overlaps in both groups (< 40 and ≥ 40 years-old) were AwObesity together with either B-Eos-high (15 and 12%, respectively) or B-Eos&FeNO-low asthma (13 and 11%) (Fig. 2).

Moreover, the proportions of subjects having AwObesity together with other phenotypes were high: 53% for the B-Eos-high phenotype, 48% for AwCOPD, 45% for the B-Eos&FeNO-low, and 44% for the FeNO-high phenotype. Also, the proportion of individuals having AwCOPD together with the B-Eos-high phenotype was high (36%), whereas the proportions were lower for the B-Eos&FeNO-low and the FeNO-high asthma phenotypes (15 and 10%, respectively) (data not shown).

In this population, only 12 and 15% of asthma subjects (< 40 and ≥ 40 years-old, respectively) with high B-Eos count had a concomitant high FeNO values (Fig. 2). Moreover, the two biomarkers were non-congruent across cut-offs. For example, when comparing groups with B-Eos count < 150/mm^3 and 150–300/mm^3, the proportion of asthma subjects having low FeNO (< 20 ppb), was not significantly different (Additional file 2: Table S1).

Associations between asthma-related outcomes and phenotype overlap

A comparison of the clinical characteristics of participants with one, two or three or more asthma phenotypes, stratified by age, is presented in Table 3 and no significant differences were observed in any age groups with regard to asthma attacks, asthma-related ED, ≥ 2 asthma symptoms, and use of ≥ 1 reliever. In the older group, the proportion of individuals with work/school absenteeism, ≥ 2 controller medications and with FEV_1/FVC < LLN was significantly higher in participants with concomitant phenotypes than in those with a single phenotype (Table 3). In both age groups, the proportion of patients with FEV_1 < LLN was significantly higher when participants presented multiple phenotypes, as well as they presented lower median FEV_1% predicted values.

When analyzing the asthma-related outcomes in subjects with a single phenotype with those having specific combination of asthma phenotypes, the overall findings were that subjects having multiple phenotypes had significantly higher proportion of using ≥ 1 reliever and ≥ 2

Table 1 Characteristics of adults with current asthma: included and excluded from phenotype classification, and stratified by single or multiple phenotypes

Demographic characteristics n (wt%)	Included subjects n = 634	Excluded subjects n = 425	p value*	Single phenotype† n = 271	Multiple phenotypes† n = 286	p value*
Female gender	410 (63)	261 (64)	0.93	174 (64)	192 (66)	0.68
Age (yrs), median (Q1–Q3)	44.0 (31.0–57.0)	48.9 (33.7–68.0)	*< 0.001*	42.0 (30.0–55.0)	47.5 (34.0–60.0)	*0.003*
BMI (kg/m^2), median (Q1–Q3)	30.8 (25.4-35.9)	31.4 (24.4-35.7)	0.99	28.7 (24.2–35.0)	33.7 (30.7–39.0)	*< 0.001*
Obesity status						
Underweight (≤ 18.4 kg/m^2)	2 (0.3)	13 (4)	*< 0.001*	1 (0.2)	1 (0.5)	0.41
Normal (18.5–24.9 kg/m^2)	142 (25)	98 (27)	0.44	78 (30)	20 (8)	*< 0.001*
Overweigh (25–29.9 kg/m^2)	151 (26)	77 (19)	0.07	80 (32)	38 (15)	*< 0.001*
Obese (≥ 30 kg/m^2)	339 (49)	187 (49)	0.94	112 (38)	227 (77)	*< 0.001*
Race and/or ethnicity						
Hispanic	105 (8)	85 (10)	0.30	38 (8)	56 (9)	0.24
Non-Hispanic white	323 (74)	184 (63)	*0.005*	135 (72)	138 (72)	0.93
Non-Hispanic black	167 (14)	117 (18)	*0.03*	76 (14)	81 (16)	0.71
Other Race	39 (4)	39 (7)	0.06	22 (6)	11 (3)	0.11
Smoking status						
Current smoker	199 (29)	114 (32)	0.40	89 (31)	87 (27)	0.46
Ex-smoker	163 (29)	98 (23)	0.13	56 (20)	89 (36)	*0.003*
Non-smoker	272 (43)	154 (45)	0.53	126 (49)	110 (37)	*0.02*
Education						
≥ High school	478 (84)	225 (69)	*< 0.001*	205 (86)	209 (81)	0.13
Asthma-related medication‡						
Reliever medication**	276 (41)	202 (48)	0.16	113 (37)	132 (47)	0.10
Oral corticosteroids	33 (8)	29 (3)	*0.001*	12 (3)	15 (4)	0.75
Inhaled corticosteroids§	153 (25)	122 (30)	0.14	55 (19)	81 (32)	*0.01*
Other control medications‖	53 (9)	63 (14)	0.19	15 (6)	33 (10)	0.50
Asthma phenotype						
AwObesity	339 (49)	–		112 (38)	227 (76)	*< 0.001*
B-Eos-high	237 (36)	–		61 (22)	176 (62)	*< 0.001*
B-Eos&FeNO-low	157 (26)	–		74 (30)	83 (30)	0.95
FeNO-high	110 (18)	–		18 (8)	92 (34)	*< 0.001*
AwCOPD	57 (8)	–		6 (2)	51 (17)	*< 0.001*
Non-classified††	77 (14)	–		–	–	

Data presented as absolute numbers and proportions weighted for the U.S. population. *p* values < 0.05 are presented in italic

Yrs years, *BMI* body mass index, *Q1* first quartile, *Q3* third quartile, *BMI* body mass index, *AwObesity* Asthma with obesity, *AwCOPD* Asthma with concurrent COPD

* Chi square test or Mann–Whitney U-test was used

† Seventy seven subjects included in the "non-classified" group were considered as missing

‡ Prescribed medication taken in the past 30 days

§ Alone or in combination with long-acting inhaled β_2-agonist

‖ Included long-acting inhaled β_2-agonist (without corticosteroids), leukotriene inhibitors, and mast cell stabilizers

**Short-acting β_2-agonist and/or anticholinergic

†† Subjects with non-single and non-multiple asthma phenotype

controller medications and had decreased lung function, with the exception of those with the B-Eos&FeNO-low phenotype combined with any of the other phenotypes (Additional file 3: Table S2, and Additional file 4: Table S3).

Moreover, a lower proportion of subjects reporting asthma attacks was observed in subjects with AwObesity and either FeNO-high (26%) or AwCOPD (20%), compared to those with a single phenotype (67%) (Additional file 3: Table S2). Subjects with concomitant AwCOPD and B-Eos&FeNO-low phenotypes had the lowest

Table 2 Demographic and clinical characteristics among all 5 phenotypes and in the "Non-classified" group

Characteristics n (wt%)	AwObesity n = 339	B-Eos-high n = 237	B-Eos&FeNO-low n = 157	FeNO-high n = 110	AwCOPD n = 57	Non-classified†† n = 77
Female gender	366 (67)	196 (55)	138 (78)	78 (52)	71 (58)	44 (56)
Age (yrs), median (Q1–Q3)	48.0 (34.0–59.0)	47.0 (31.0–59.0)	41.0 (27.0–57.0)	45.0 (30.0–54.0)	61.0 (52.0–69.0)	39.0 (28.0–53.0)
BMI (kg/m^2), median (Q1–Q3)	35.4 (32.5–40.4)	30.3 (25.9–36.8)	28.9 (24.2–33.0)	27.6 (24.9–33.4)	30.3 (25.1–35.1)	24.3 (22.8–27.5)
Race and/or ethnicity						
Hispanic	94 (9)	78 (11)	30 (8)	29 (10)	18 (4)	11 (6)
Non-Hispanic white	230 (65)	175 (73)	79 (69)	65 (75)	80 (81)	50 (86)
Non-Hispanic black	177 (21)	74 (12)	62 (17)	38 (13)	27 (11)	10 (6)
Other race	25 (5)	21 (4)	13 (6)	6 (2)	10 (4)	6 (2)
Smoking status						
Current smoker	150 (27)	104 (32)	48 (23)	18 (13)	65 (52)	23 (28)
Ex-smoker	134 (27)	100 (32)	35 (25)	42 (39)	70 (48)	18 (30)
Non-smoker	229 (45)	124 (35)	88 (51)	66 (49)	0 (0)	36 (42)
Education						
≥ High school	358 (79)	227 (80)	125 (81)	102 (87)	73 (59)	64 (91)
Asthma-related medication‡						
Reliever medication**	234 (42)	172 (51)	82 (44)	65 (46)	69 (57)	31 (36)
Oral corticosteroids	36 (5)	17 (5)	4 (2)	10 (7)	15 (7)	2 (2)
Inhaled corticosteroids§	144 (27)	103 (32)	36 (20)	36 (33)	60 (47)	17 (19)
Asthma-related outcomes						
Asthma attack	363 (68)	252 (74)	125 (68)	98 (71)	85 (63)	54 (75)
Asthma-related ED	130 (27)	80 (23)	41 (23)	26 (13)	37 (32)	9 (8)
> 2 asthma symptoms	311 (66)	199 (65)	92 (55)	80 (57)	90 (74)	42 (59)
Work/school absenteeism	66 (18)	43 (16)	25 (18)	23 (14)	12 (20)	10 (14)
Lung function						
FEV_1% predicted, median (Q1–Q3)	89.0 (75.6–99.2)	84.1 (75.2–95.4)	93.2 (83.8–100.8)	82.7 (75.9–95.4)	74.0 (62.9–90.1)	89.9 (80.5–103.5)
FEV_1/FVC, median (Q1–Q3)	0.77 (0.62–0.82)	0.74 (0.65–0.80)	0.79 (0.72–0.83)	0.72 (0.66–0.79)	0.63 (0.50–0.75)	0.76 (0.69–0.82)

Data presented as absolute numbers and proportions weighted for the U.S. population

Yrs years, *BMI* body mass index, *Q1* first quartile, *Q3* third quartile, *BMI* body mass index, *AwObesity* Asthma with obesity, *AwCOPD* Asthma with concurrent (COPD), *ED* emergency-department, *FEV1* Forced expiratory volume in 1 s, *FEV1/FVC* forced expiratory volume in 1 s and functional vital capacity ratio, *LLN* lower limit of normality

†† Subjects with non-single and non-multiple asthma phenotype

‡ Prescribed medication taken in the past 30 days

** Short-acting β_2-agonist and/or anticholinergic

§ Alone or in combination with long-acting inhaled β_2-agonist

proportion of ≥ 2 asthma symptoms (20%), but had the highest proportion of using ≥ 1 reliever medication (84%) as well as having FEV_1 < LLN (71%).

In multivariate regression analysis, adjusting for co-variables, having multiple phenotypes was significantly associated with using ≥ 2 controller medications (aOR, 95% CI 2.03, 1.16–3.57), and having reduced FEV_1 (3.21, 1.73–5.94) (Table 4). However, no associations were seen with asthma attacks, asthma-related ED, ≥ 2 asthma symptoms, work/school absenteeism, use of reliever medication or FEV_1/FVC < LLN (Additional file 5: Table S4).

Furthermore, subjects aged ≥ 40 years-old, had significantly higher odds of using ≥ 2 controller medications and having FEV_1 < LLN predicted, compared to those < 40 years-old, adjusted for covariates (Table 4). Being a current smoker was significantly associated with using ≥ 1 reliever medication (1.95, 1.35–2.83) and with reduced lung function: FEV_1 < LLN predicted (2.01, 1.21–3.33) and FEV_1/FVC < LLN (2.02, 1.16–3.51) and not associated with any other asthma-related outcomes (Table 4 and Additional file 5: Table S4).

The association between having concomitant phenotypes and using multiple controller medications was consistent when considering oral corticosteroids (OCS) separated from other controller medications (1.87, 1.09–3.21) (data not shown).

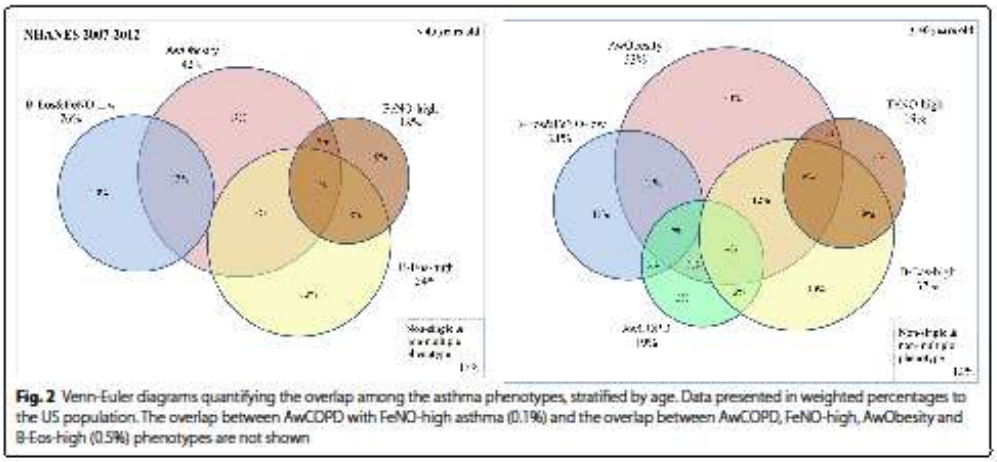


Fig. 2 Venn-Euler diagrams quantifying the overlap among the asthma phenotypes, stratified by age. Data presented in weighted percentages to the US population. The overlap between AwCOPD with FeNO-high asthma (0.1%) and the overlap between AwCOPD, FeNO-high, AwObesity and B-Eos-high (0.5%) phenotypes are not shown

We also analyzed the potential bias of controller medications in the phenotype classification, particularly in the B-Eos-high and FeNO-high phenotypes (Additional file 6: Fig. S1). No significant differences in asthma-related treatment were found between the phenotypes, with exception for a higher proportion of patients treated with ICS within the FeNO-high and B-Eos-high phenotypes compared to those with B-Eos&FeNO-low phenotype ($p = 0.03$). When restricting to subjects with a single asthma phenotype no significant differences were found.

Moreover, sensitivity analyses showed that the proportion of total overlap (weighted 53%), and the associations between having multiple phenotypes and asthma outcomes were similar when imputing all missing values (data not shown). The goodness-of-fit test revealed adequate fitting for all regression models, except when using $FEV_1/FVC < LLN$ as dependent variable (Additional file 5: Table S4) and no statistically significant interactions between co-variables were observed.

Discussion

We report a substantial overlap of commonly reported asthma phenotypes among adults with current asthma in a large population sample, with almost half of them having two or more concomitant phenotypes. Furthermore, having multiple asthma phenotypes, regardless of their classification, was associated with poorer asthma outcomes, particularly the use of more controller medication and reduced lung function.

These findings illustrate the complexity and unique features of the concomitant asthma phenotypes when categorizing asthma in adults, using only the "classical" (hypothesis-driven) approach, based on measures readily available in the clinic (such as non-invasive biomarkers and medical records).

Hypothesis-driven asthma phenotypes are usually based on single dimensions of the disease, such as clinical symptoms, triggers, pathology or patterns of airway obstruction [16–18, 40–42]. However, evidence has shown that this approach is highly heterogenous, as it depends on the a priori assumptions and target population [43–45]. Also, it is of note that the 77 subjects with asthma that could not be classified as having any of the studied phenotypes, supports the fact that there is a considerable number of asthma patients whose clinical phenotype is not easily classified (e.g. asthmatics with irreversible airflow obstruction, patients with similar airways symptoms but with different pattern of airway inflammation), suggesting the presence of sub-phenotypes [1, 22, 44, 45].

In an attempt to explore the pathophysiology of specific asthma subgroups, and help stratify patients for targeted therapies, data-driven or unsupervised approaches (such as k-means, hierarchical clustering, partition-around-medoids methods or latent class analysis) are being applied in airways disease to identify "novel" accurate and distinct phenotypes, taking into account the heterogeneity and multidimensional characteristics of the disease [8, 46–52].

Our study results seem to be in line with the view of those that argue for a combination of both hypothesis- and data-driven approaches as a way forward to progress our knowledge on asthma endotypes and clinical phenotypes in an iterative way [52–54]. The data-driven

Table 3 Distribution and comparisons of the asthma-related outcomes among asthma phenotypes, stratified by age

	Total n (wt%)	Asthma attack	Asthma-related ED	≥ 2 asthma symptoms	Work/school absenteeism	Asthma medication		Lung function		
						≥ 1 reliever medication[†]	≥ 2 controller medication	FEV_1 <LLN	FEV_1% predicted[§]	FEV_1/FVC <LLN
< 40 yrs										
1 phenotype	118 (56)	85 (70)	27 (20)	59 (96)	15 (21)	55 (43)	17 (11)	10 (7)	95.6 (90.2–102.1)	26 (20)
2 phenotypes	97 (40)	69 (75)	26 (32)	57 (73)	12 (9)	41 (40)	8 (9)	19 (23)	91.8 (81.3–99.2)	25 (29)
3 phenotypes	12 (4)	10 (73)	5 (40)	8 (72)	4 (25)	9 (68)	1 (4)	3 (36)	81.9 (75.3–84.6)	4 (29)
p value*										
1 versus 2		0.43	0.10	0.052	*0.04*	0.72	0.64	*0.006*	0.07	0.25
2 versus 3		0.92	0.55	0.96	*0.03*	0.15	0.46	0.43	*0.009*	0.98
1 versus 3		0.86	0.11	0.39	0.73	0.20	0.34	*0.01*	*<0.001*	0.48
≥ 40 yrs										
1 phenotype	153 (46)	104 (65)	35 (23)	76 (62)	20 (15)	58 (33)	26 (17)	29 (20)	91.2 (81.6–99.2)	34 (26)
2 phenotypes	136 (41)	92 (69)	28 (21)	70 (55)	12 (9)	60 (46)	43 (36)	31 (37)	80.4 (70.0–91.7)	27 (32)
≥ 3 phenotypes	41 (13)	26 (72)	4 (9)	30 (70)	8 (40)	22 (60)	16 (40)	12 (46)	74.0 (63.0–85.8)	16 (53)
p value*										
1 versus 2		0.56	0.82	0.38	0.41	0.12	*0.02*	*0.01*	*0.007*	0.46
2 versus ≥ 3		0.80	0.27	0.30	*0.002*	0.24	0.67	0.47	0.22	0.09
1 versus ≥ 3		0.50	0.21	0.54	*0.03*	*0.053*	*0.02*	*0.01*	*0.006*	*0.02*

Data presented as absolute numbers and proportions weighted for the U.S. population. *p* values < 0.05 are presented in italic. The 77 subjects included in the "non-classified" group were considered as missing

ED emergency department, FEV_1 forced expiratory volume in 1 s, FEV_1/FVC forced expiratory volume in 1 s and functional vital capacity ratio, *LLN* lower limit of normality, Q1 first quartile; Q3: third quartile

* Chi square test or Mann–Whitney U-test was used

† Short-acting β_2-agonist or/and anticholinergic

§ Presented as median (Q1–Q3)

Table 4 Regression models with significant associations between having multiple asthma phenotypes and asthma-related outcomes, adjusted for co-variates

	≥ 2 controller medications		FEV_1 < LLN	
	aOR	95% CI	aOR	95% CI
Multiple versus single phenotype	*2.03*	*1.16–3.57*	*3.21*	*1.74–5.94*
Female gender	1.39	0.77–2.50	1.51	0.81–2.81
Age ≥ 40 yrs	*3.01*	*1.52–5.95*	*2.55*	*1.29–5.03*
Caucasian versus others	1.38	0.86–2.23	1.37	0.78–2.42
Current smoker versus non-/ex-smokers	1.02	0.52–2.02	*2.01*	*1.21–3.33*
Rhinitis	1.08	0.57–2.16	0.94	0.54–1.63
Goodness-of-fit test				
χ^2 (*p* value)	0.86 (0.56)		0.80 (0.61)	

Multivariate logistic regression models adjusted for gender, age, race, current smoking and rhinitis. The aOR values with $p < 0.05$ are presented in Italic

FEV_1 forced expiratory volume in 1 s, *LLN* lower limit of normal, *CI* confidence interval, *aOR* adjusted odds ratio, χ^2 Chi square goodness-of-fit

phenotypes studies obtained some of the phenotypes already defined by hypothesis-driven approaches (e.g. obese, non-eosinophilic asthmatics [8]; persistent airway inflammation [46]; low type-2 inflammation [49]; fixed obstructive, non-eosinophilic and neutrophilic [50]), but, importantly, they identified other phenotypes that differ by certain characteristics: clinical parameters [8, 47–49], clinical response to treatment [46, 52], and airway inflammation [49, 51]. Therefore, further studies are required to compare and validate the asthma phenotypes obtained using different unsupervised methods.

The high overlap of asthma phenotypes seen in this study was similar to the findings of Tran et al. [15], who used datasets from previous NHANES surveys to evaluate the overlap of asthma phenotypes. However, Tran et al. study focused on allergic asthma phenotypes, based on IgE levels, and was therefore limited to the 2005–2006 survey that lacks data on FeNO. We provided a broader analysis of phenotypes that included not only the eosinophil-based phenotype (associated more closely with IL-5-driven) and the one based on FeNO values (mostly dependent on IL-4/IL-13-driven) [12], but also other phenotypes not defined by biomarkers and in a much larger dataset.

In this study, we extended previous observations [7, 11] suggesting that FeNO and B-Eos count partially reflect different inflammatory pathways, representing a local and a systemic type 2-marker, respectively. We observed that only 12–15% of asthma subjects with high B-Eos count had concomitant high FeNO levels, in this population. Also, a similar proportion of subjects with multiple phenotypes was obtained when considering the "Type 2-high" phenotype, supporting the view that these two biomarkers are not interchangeable and that the use of both biomarkers in combination may allow for better targeted and personalized treatment for at least certain subsets of asthma patients [7, 10, 11].

The more prevalent combinations of phenotypes observed in this study were AwObesity together with either B-Eos-high or B-Eos&FeNO-low phenotypes. This supports the view that obesity-related asthma, despite often suggested to be a separate asthma phenotype associated with non-eosinophilic airway inflammation [9, 55, 56], may also be associated to eosinophilic inflammation.

Given the high prevalence in the US population, in this sample, obesity is likely to be a comorbidity, rather than the primary reason for asthma [57]; however, we defined the AwObesity phenotype as a separate group, since the interdependence on inflammatory markers to targeting different asthma therapies makes essential the accurate characterization of inflammation in obese asthmatic subjects [19, 58]. In addition, the relevance of defining the AwObesity phenotype is supported by the data as the weighted proportion of overlap is similar when excluding AwObesity or B-Eos-high phenotype from the analysis (data not shown). Moreover, the weighted proportion of subjects with "non-classified" asthma doubled when excluding the AwObesity phenotype (increasing to 32% in the < 40 years-old and 29% in ≥ 40 years old).

Having multiple asthma phenotypes was more common in older subjects and in non- and former- smokers; whether this is due to a general increase of comorbidities with age [58–60] and/or an interaction with environmental factors [61] cannot be specifically addressed by this study design. Nevertheless, when interpreting these results one should bear in mind that AwCOPD is associated to older age and prior/current smoking while FeNO increases with age and decreases with smoking [62, 63].

Interestingly, subjects with a higher number of concomitant asthma phenotypes presented reduced lung function and this association remained when controlling for potential confounders by multiple regression analysis. This shows that having multiple phenotypes is independently associated with reduced lung function, suggesting a cumulative effect of different disease processes.

Moreover, patients having several commonly reported asthma phenotypes had higher odds of using more controller medications, supporting the view that these patients are those with more complex disease and higher asthma morbidity [9, 15]. This also suggests that these asthma patients have an inadequate response to prescribed therapies, since lung function was reduced, and that they may represent a group of patients with the need for add-on treatment. However, the choice of specific

treatments, such as biological therapies, will be more difficult considering the complexity introduced by having multiple phenotypes.

The lack of significant associations between multiple phenotypes and the other asthma-related outcomes may be difficult to understand. A possible explanation could be that the prescribed medication is effective against asthma symptoms and attacks but less effective against (subclinical) processes that cause long-term reduction in lung function. However, the results could also be related to data collection methods, as lung function measurement and the way medication use was ascertained, were less dependent on patient recall than the self-reported variables that were used for the outcomes with null results in the present study. Further studies should be done, adding the quantitative assessment of asthma attacks/asthma-related ED visits, and also including the age of asthma onset, that could have an influence on asthma-related outcomes.

This study has several limitations. First, because of its cross-sectional design, it was not possible to evaluate interactions between phenotypes over time in patients with concomitant phenotypes, nor was it possible to determine which phenotype occurred first. Second, although there were differences between the included and excluded groups in the variables age, BMI, non-Hispanic white/black subjects, having finished high school and OCS use, the majority were not used for phenotypic classification and did not affect the outcomes, as shown in sensitivity analysis. Third, as our asthma and COPD definitions were based on self-reported diagnosis, rather than relying on lung function tests, the acquired information is subject to recall bias and misclassification. However, these definitions have been commonly used in NHANES reports [7, 11, 15] and have proven to be reasonably reliable [64, 65]. Moreover, we have used the most frequent combination of questions seen in epidemiological studies [65, 66], and we also included questions on recent wheeze and/or asthma attacks, which should reduce the risk of including individuals without true disease. Also, we stratified the analysis by age (at 40 years), as used in other COPD studies [37, 67], in order to improve the clinical value into the interpretation of phenotyping data, as the overlap among asthma phenotypes will be different among those less than age 40 and those older than 40 (with higher possibility of having COPD) [37]. Fourth, the lack of other biomarkers in the present NHANES years, prevented the analysis of other asthma phenotypes and the use of alternative definitions. However, we analyzed biomarkers of type 2-inflammation in both blood and exhaled air that previously have been shown to independently relate to asthma morbidity [7, 11]. Fifth, as there is no consensual definition of biomarker-defined asthma phenotypes, we based our definitions on cut-offs used in previous studies to discriminate patients in single asthma phenotypes [32–35], rather than on any reported specificity or sensitivity for predicting asthma morbidity or response to therapeutics [68, 69]. For high probability of eosinophilic inflammation, the cut-off value for FeNO has been suggested to be >50 ppb for adults [70]. However, we chose a FeNO cut-off of 35 ppb, based on the mean baseline FeNO levels of patients included in randomized controlled trials of anti-IL-13 treatment [33, 69]. In spite of using this lower cut-off, 77 subjects with current asthma could not be classified as having any of the predefined phenotypes, indicating the need for better, and probably personalized, cut-offs for biomarkers in asthma.

Furthermore, we could not demonstrate a clear effect of ongoing controller medication in the phenotypic classification in our data. No significant associations were observed, probably at least partially explained with the exclusion of the participants with ICS/OCS use <48 h prior to the exhaled NO measurements. Also, contrary to the expected, we observed a higher proportion of patients treated with ICS/OCS within both B-Eos-high and FeNO-high phenotypes than the B-Eos&FeNO-low phenotype. A plausible explanation is that subjects with ongoing inflammation have more clear asthma, with more symptoms, and, thus, a higher need of treatment. B-Eos&FeNO-low asthma is a heterogeneous group with less need of controller treatment, and because the treatment is ineffective, it may be that medication use and even prescription has been stopped.

Finally, even though we did not specifically analyze the overlap of asthma phenotypes in patients with severe asthma, the significant association between presenting more than one phenotype and being treated with multiple asthma controller medications suggests higher asthma severity in this subset [71]. Also, we did not consider individual environmental factors, such as air pollution and/or indoor allergens, that could influence asthma phenotypes. Further studies describing the overlap in patients with severe asthma and studies examining asthma patients exposed to different environmental factors, such as subjects who live in cities versus in rural areas are needed.

This study indicates that the overlap of commonly reported asthma phenotypes is observed also in non-selected asthma patients from the general population. Our findings highlight the importance of classifying asthma patients with regard to applicable phenotypes, rather than using a single asthma phenotype, to enable the development of adequate targeted strategies to avoid lung function impairment. However, further data is required, such as that from higher order analysis, using

data-mining methods possibly combined with those that rely on predefined hypotheses. This synergy is expected to improve the knowledge on asthma phenotypes and, ultimately, to lead to more personalized treatment strategies [53, 54].

Conclusions

In conclusion, a prevalent overlap of commonly reported phenotypes was observed in asthma patients identified from the general population. Subjects classified as having multiple phenotypes used more controller medications and had reduced lung function. Thus, the complexity and unique features of concomitant asthma phenotypes may require a broader data analysis approach, based on a combination of clinical information and biomarkers resulting in better characterization of patients. This could lead to better asthma outcomes, particularly preserved lung function.

Additional files

Additional file 1. Supplementary Methods.

Additional file 2: Table S1. Distribution and comparisons between the FeNO and B-Eos cut-offs used in this study, among individuals with current asthma.

Additional file 3: Table S2. Weighted percentages and comparisons of asthma-related outcomes among subjects with a single asthma phenotype versus: non-classified, and specific combinations of asthma phenotypes.

Additional file 4: Table S3. Weighted percentages and comparisons of asthma-related outcomes among subjects with a single asthma phenotype versus: non-classified, and specific combinations of asthma phenotypes.

Additional file 5: Table S4. Multivariable logistic regression models between each asthma-related outcome and having multiple asthma phenotypes, adjusted for co-variables.

Additional file 6: Fig. S1. Proportions (weighted to the US population) of subjects taking asthma controller medications stratified into the different phenotypes, among all participants included for asthma phenotype classification (left) and only in those with a single phenotype (right). P-values <0.05 were indicated. NA: Non-applicable (not possible to determine because some participants had both B-Eos-high and FeNO-high asthma phenotypes).

Abbreviations

aOR: adjusted odds ratios; ATS/ERS: American Thoracic Society/European Respiratory Society; AwCOPD: asthma with concurrent COPD; AwObesity: asthma with obesity; B-Eos: blood eosinophils; BMI: body mass index; COPD: chronic obstructive pulmonary disease; CI: confidence intervals; ED: emergency department; FeNO: fraction of exhaled nitric oxide; FEV_1: forced expiratory volume in 1 s; FEV_1/FVC: ratio of forced expiratory volume in 1 s to forced vital capacity; ICS: inhaled corticosteroids; IgE: immunoglobulin E; IL: interleukin; LLN: lower limit of normal; NHANES: National Health and Nutrition Examination Survey (USA); OCS: oral corticosteroids; Q1: first quartile; Q3: third quartile; Th2: T helper cell type 2.

Authors' contributions

RA, JAF, AM and KA contributed to study conception and design, analysis and interpretation of data, writing and revising the article. TJ, AMP and CJ contributed to data interpretation, writing and revising the article. All authors read and approved the final version of the manuscript.

Author details

[1] CINTESIS- Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Edifício Nascente, Piso 2, Rua Dr. Plácido da Costa, s/n, 4200-450 Porto, Portugal. [2] Department of Cardiovascular and Respiratory Sciences, Porto Health School, Porto, Portugal. [3] MEDCIDS- Department of Community Medicine, Information, and Health Sciences, Faculty of Medicine, University of Porto, Porto, Portugal. [4] Department of Allergy, Instituto & Hospital CUF, Porto, Portugal. [5] Department of Medical Sciences: Clinical Physiology, Uppsala University, Uppsala, Sweden. [6] Department of Medical Sciences: Respiratory Medicine and Allergology, Uppsala University, Uppsala, Sweden. [7] Department of Women's and Children's Health: Paediatric Research, Uppsala University, Uppsala, Sweden.

Acknowledgements

RA is supported by Fundação para a Ciência e Tecnologia (PhD grant – Ref. PD/BD/113659/2015), Portugal.

Competing interests

The authors declare that they have no competing interests.

Availability of data

The datasets used and analyzed during this study are available in the NHANES website: https://www.cdc.gov/nchs/nhanes/index.htm.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The NHANES survey operates under the approval of the National Center for Health Statistics Research Ethics Review Board, Protocols #2005-06, and #2011-17 (www.cdc.gov/nchs/nhanes/irba98.htm). All of the NHANES data meet the circumstances described in Policy #39, Research Using Publicly Available Datasets (Secondary Analysis) for use without application to Institutional Review Board. Moreover, all study participants provided written informed consent.

Funding

Project "NORTE-01-0145-FEDER-000016" is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 12 December 2017 **Accepted:** 26 March 2018
Published online: 30 May 2018

References

1. Froidure A, Mouthuy J, Durham SR, Chanez P, Sibille Y, Pilette C. Asthma phenotypes and IgE responses. Eur Respir J. 2016;47:304–19.
2. Corren J. Asthma phenotypes and endotypes: an evolving paradigm for classification. Discov Med. 2013;15(83):243–9.
3. Woodruff PG, Modrek B, Choy DF, Jia G, Abbas AR, Ellwanger A, Koth LL, Arron JR, Fahy JV. T-helper type 2-driven inflammation defines major subphenotypes of asthma. Am J Respir Crit Care Med. 2009;180(5):388–95.
4. Robinson D, Humbert M, Buhl R, Cruz AC, Inoue H, Korom S, Hanania NA, Nair P. Revisiting Type 2-high and Type 2-low airway inflammation in asthma: current knowledge and therapeutic implications. Clin Exp Allergy. 2017;47(2):161–75.
5. Hastie AT, Moore WC, Li H, Rector BM, Ortega VE, Pascual RM, Peters SP, Meyers DA, Bleecker ER, National Heart, Lung, and Blood Institute's Severe Asthma Research Program. Biomarker surrogates do not accurately

predict sputum eosinophil and neutrophil percentages in asthmatic subjects. J Allergy Clin Immunol. 2013;132(1):72.e12–80.e12.
6. Green RH, Brightling CE, Woltmann G, Parker D, Wardlaw AJ, Pavord ID. Analysis of induced sputum in adults with asthma: identification of subgroup with isolated sputum neutrophilia and poor response to inhaled corticosteroids. Thorax. 2002;57(10):875–9.
7. Malinovschi A, Fonseca JA, Jacinto T, Alving K, Janson C. Exhaled nitric oxide levels and blood eosinophil counts independently associate with wheeze and asthma events in National Health and Nutrition Examination Survey subjects. J Allergy Clin Immunol. 2013;132(4):821.e5–27.e5.
8. Jia G, Erickson RW, Choy DF, Mosesova S, Wu LC, Solberg OD, et al. Periostin is a systemic biomarker of eosinophilic airway inflammation in asthmatic patients. J Allergy Clin Immunol. 2012;130(3):647.e10–54.e10.
9. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med. 2008;178(3):218–24.
10. Jia G, Erickson RW, Choy DF, Mosesova S, Wu LC, Solberg OD, Shikotra A, Carter R, Audusseau S, Hamid Q, et al. Identification of airway mucosal type 2 inflammation by using clinical biomarkers in asthmatic patients. J Allergy Clin Immunol. 2017;140(3):710–9.
11. Malinovschi A, Janson C, Borres M, Alving K. Simultaneously increased fraction of exhaled nitric oxide levels and blood eosinophil counts relate to increased asthma morbidity. J Allergy Clin Immunol. 2016;138(5):1301.e2–308.e2.
12. Kattal RK, Bensch GW, Busse WW, Chipps BE, Denson JL, Garber AN, Jacobs JS, Kraft M, Martin RJ, Nair P, et al. Changing paradigms in the treatment of severe asthma: the role of biologic therapies. J Allergy Clin Immunol Pract. 2017;5(2):51–14.
13. Chung KF, Adcock IM. How variability in clinical phenotypes should guide research into disease mechanisms in asthma. Ann Am Thorac Soc. 2013;10(Suppl):S109–17.
14. Merritt F. Blood eosinophils: the Holy Grail for asthma phenotyping? Ann Allergy Asthma Immunol. 2016;116(2):90–1.
15. Tran TN, Zeiger RS, Peters SP, Colice G, Newbold P, Goldman M, Chipps BE. Overlap of atopic, eosinophilic, and TH2-high asthma phenotypes in a general population with current asthma. Ann Allergy Asthma Immunol. 2016;116(1):37–42.
16. Borish L. The immunology of asthma: asthma phenotypes and their implications for personalized treatment. Ann Allergy Asthma Immunol. 2016;117(2):108–14.
17. Hekking PPW, Bel EH. Developing and emerging clinical asthma phenotypes. J Allergy Clin Immunol Pract. 2014;2(6):671–80.
18. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nat Med. 2012;18(5):716–25.
19. Gibeon D, Batuwita K, Osmond M, Heaney LG, Brightling CE, Niven R, et al. Obesity-associated severe asthma represents a distinct clinical phenotype: analysis of the British Thoracic Society Difficult Asthma Registry Patient cohort according to BMI. Chest. 2013;143(2):406–14.
20. Kobayashi S, Hanagama M, Yamanda S, Ishida M, Yanai M. Inflammatory biomarkers in asthma-COPD overlap syndrome. Int J Chronic Obstr Pulm Dis. 2016;11:2117–23.
21. Carr TF, Zaki AA, Kraft M. Eosinophilic and noneosinophilic asthma. Am J Respir Crit Care Med. 2018;197(1):22–37.
22. Christenson SA, Steiling K, van den Berge M, Hijazi K, Hiemstra PS, Postma DS, Lenburg ME, Spira A, Woodruff PG, et al. Asthma-COPD overlap. Clinical relevance of genomic signatures of type 2 inflammation in chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2015;191(7):758–66.
23. Complete Blood Count. http://www.cdc.gov/nchs/data/nhanes/nhanes_11_12/cbc_g_met_he.pdf. Accessed 17 June 2017.
24. Respiratory Health ENO Procedures Manual. www.cdc.gov/nchs/data/nhanes/nhanes_11_12/Respiratory_Health_ENO_Procedures_Manual.pdf. Accessed 26 May 2017.
25. Respiratory Health Spirometry Procedures Manual. https://www.cdc.gov/nchs/data/nhanes/nhanes_11_12/spirometry_procedures_manual.pdf. Accessed 6 June 2017.
26. ATS/ERS. ATS/ERS Recommendations for Standardized Procedures for the Online and Offline Measurement of Exhaled Lower Respiratory Nitric Oxide and Nasal Nitric Oxide, 2005. Am J Respir Crit Care Med. 2005;171(8):912–30.
27. Miller MR. Standardisation of spirometry. Eur Respir J. 2005;26(2):319–38.
28. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med. 1999;159(1):179–87.
29. Hankinson JL, Kawut SM, Shahar E, Smith LJ, Stukovsky KH, Barr RG. Performance of American Thoracic Society-recommended spirometry reference values in a multiethnic sample of adults. Chest. 2010;137(1):138–45.
30. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, Coates A, van der Grinten CP, Gustafsson P, Hankinson J, et al. Interpretative strategies for lung function tests. Eur Respir J. 2005;26(5):948–68.
31. Prescription Medications - Drug Information. http://wwwn.cdc.gov/Nchs/Nhanes/1999-2000/RXQ_DRUG.htm. Accessed 6 June 2017.
32. Castro M, Zangrilli J, Wechsler ME, Bateman ED, Brusselle GG, Bardin P, Murphy K, Maspero JF, O'Brien C, Korn S. Reslizumab for inadequately controlled asthma with elevated blood eosinophil counts: results from two multicentre, parallel, double-blind, randomised, placebo-controlled, phase 3 trials. Lancet Respir Med. 2015;3:355–66.
33. Wenzel S, Ford L, Pearlman D, Spector S, Sher L, Skobieranda F, Wang L, Kirkesseli S, Rocklin R, Bock B, et al. Dupilumab in persistent asthma with elevated eosinophil levels. N Engl J Med. 2013;368(26):2455–66.
34. Dweik RA, Sorkness RL, Wenzel S, Hammel J, Curran-Everett D, Comhair SA, Bleecker E, Busse W, Calhoun WJ, Castro M, et al. Use of exhaled nitric oxide measurement to identify a reactive, at-risk phenotype among patients with asthma. Am J Respir Crit Care Med. 2010;181(10):1033–41.
35. McGrath KW, Icitovic N, Boushey HA, Lazarus SC, Sutherland ER, Chinchilli VM, Fahy JV, Asthma Clinical Research Network of the National Heart, Lung, and Blood Institute. A large subgroup of mild-to-moderate asthma is persistently noneosinophilic. Am J Respir Crit Care Med. 2012;185(6):612–9.
36. Flegal KM, Carroll MD, Kuczmarski RJ, Johnson CL. Overweight and obesity in the United States: prevalence and trends, 1960–1994. Int J Obes Relat Metab Disord. 1998;22(1):39–47.
37. Buist AS, McBurnie MA, Vollmer WM, Gillespie S, Burney P, Mannino DM, et al. International variation in the prevalence of COPD (The BOLD Study): a population-based prevalence study. Lancet. 2007;370(9589):741–50.
38. Diagnosis of Diseases of Chronic Airflow Limitation: Asthma, COPD and Asthma-COPD Overlap Syndrome (ACOS). http://goldcopd.org/asthma-copd-asthma-copd-overlap-syndrome/. Accessed 3 Jan 2017.
39. Archer KJ, Lemeshow S. Goodness-of-fit test for a logistic regression model fitted using survey sample data. Stata J. 2006;6(1):97–105.
40. Hirano T, Matsunaga K. Late-onset asthma: current perspectives. J Asthma and Allergy. 2018;11:19–27.
41. Wenzel SE. Asthma: defining of the persistent adult phenotypes. Lancet. 2006;368:804–13.
42. Vonk JM, Jongepier H, Panhuysen CIM, Schouten JP, Bleecker ER, Postma DS. Risk factors associated with the presence of irreversible airflow limitation and reduced transfer coefficient in patients with asthma after 26 years of follow up. Thorax. 2003;58(4):322–7.
43. Bousquet J, Anto JM, Sterk PJ, Adcock IM, Chung KF, Roca J, et al. Systems medicine and integrated care to combat chronic noncommunicable diseases. Genome Med. 2011;3:43.
44. Agusti A, Bel E, Thomas M, Vogelmeier C, Brusselle G, Holgate S, Humbert M, Jones P, Gibson PG, Vestbo J, et al. Treatable traits: toward precision medicine of chronic airway diseases. Eur Respir J. 2016;47(2):410–9.
45. Zedan MM, Laimon WN, Osman AM, Zedan MM. Clinical asthma phenotyping: a trial for bridging gaps in asthma management. World J Clin Pediatr. 2015;4(2):13–8.
46. Wu W, Bleecker E, Moore W, Busse WW, Castro M, Chung KF, Calhoun WJ, Erzurum S, Gaston B, Israel E, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. J Allergy Clin Immunol. 2014;133(5):1280–8.
47. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R Jr, Castro M, Curran-Everett D, Fitzpatrick AM, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med. 2010;181(4):315–23.
48. Siroux V, Basagaña X, Boudier A, Pin I, Garcia-Aymerich J, Vesin A, Slama R, Jarvis D, Anto JM, Kauffmann F, Sunyer J. Identifying adult asthma phenotypes using a clustering approach. Eur Respir J. 2011;38(2):310–7.
49. Lefaudeux D, de Meulder B, Loza MJ, Peffer N, Rowe A, Baribaud F, Bansal AT, Lutter R, Sousa AR, Corfield J, et al. U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. J Allergy Clin Immunol. 2017;139:1797–807.

50. Loza MJ, Adcock I, Auffray C, Chung KF, Djukanovic R, Sterk PJ, Susulic VS, Barnathan ES, Baribaud F, Silkoff PE, et al. Longitudinally stable, clinically defined clusters of patients with asthma independently identified in the ADEPT and U-BIOPRED asthma studies. Ann Am Thorac Soc. 2016;13(Suppl 1):S102–3.
51. Loza MJ, Djukanovic R, Chung KF, Horowitz D, Ma K, Branigan P, Barnathan ES, Susulic VS, Silkoff PE, Sterk PJ, et al. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. Respir Res. 2016;17(1):165.
52. Clemmer GL, Wu AC, Rosner B, McGeachie MJ, Litonjua AA, Tantisira KG, Weiss ST. Measuring the corticosteroid responsiveness endophenotype in asthmatic patients. J Allergy Clin Immunol. 2015;136(2):274.e8–81.e8.
53. Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, Custovic A. Disaggregating asthma: big investigation vs. big data. J Allergy Clin Immunol. 2016;139(2):400–7.
54. Bousquet J, Anto JM, Akdis M, Auffray C, Keil T, Momas I, Postma DS, Valenta R, Wickman M, Cambon-Thomsen A, et al. Paving the way of systems biology and precision medicine in allergic diseases: the MeDALL success story. Allergy. 2016;71(11):1513–25.
55. Leiria LOS, Martins MA, Saad MJA. Obesity and asthma: beyond TH2 inflammation. Metabolism. 2015;64(2):172–81.
56. Bates JHT, Poynter ME, Frodella CM, Peters U, Dixon AE, Suratt BT. Pathophysiology to phenotype in the asthma of obesity. Ann Am Thorac Soc. 2017;14(Supplement 5):S395–8.
57. Gonzalez-Barcala FJ, Pertega S, Perez-Castro T, Sampedro M, Sanchez-Lastres J, San-Jose-Gonzalez MA, Bamonde L, Garnelo L, Valdes-Cuadrado L, Moure JD, et al. Obesity and asthma: an association modified by age. Allergol Immunopathol (Madr). 2013;41(3):176–80.
58. Amelink M, De Nijs SB, De Groot JC, Van Tilburg PMB, Van Spiegel PI, Krouwels FH, et al. Three phenotypes of adult-onset asthma. Allergy. 2013;68(5):674–80.
59. Ledford DK, Lockey RF. Asthma and comorbidities. Curr Opin Allergy Clin Immunol. 2013;13(1):78–86.
60. Yanez A, Cho S-H, Soriano JB, Rosenwasser LJ, Rodrigo GJ, Rabe KF, Peters S, Nilmi A, Ledford DK, Kattal R, et al. Asthma in the elderly: what we know and what we have yet to know. World Allergy Organ J. 2014;7(1):8.
61. Miller RL, Ho SM. Environmental epigenetics and asthma: current concepts and call for studies. Am J Respir Crit Care Med. 2008;177(6):567–73.
62. Jacinto T, Malinovschi A, Janson C, Fonseca J, Alving K. Evolution of exhaled nitric oxide levels throughout development and aging of healthy humans. J Breath Res. 2015;9(3):36005.
63. Jacinto T, Malinovschi A, Janson C, Fonseca J, Alving K. Differential effect of cigarette smoke exposure on exhaled nitric oxide and blood eosinophils in healthy and asthmatic individuals. J Breath Res. 2017;11(3):036006.
64. Hallidin CN, Doney BC, Hnizdo E. Changes in prevalence of chronic obstructive pulmonary disease and asthma in the US population and associated risk factors. Chron Respir Dis. 2015;12(1):47–60.
65. Sá-Sousa A, Jacinto T, Azevedo LF, Morais-Almeida M, Robalo-Cordeiro C, Bugalho-Almeida A, Bousquet J, Fonseca JA. Operational definitions of asthma in recent epidemiological studies are inconsistent. Clin Transl Allergy. 2014;4:24.
66. de Marco R, Cerveri I, Bugiani M, Ferrari M, Verlato G. An undetected burden of asthma in Italy: the relationship between clinical and epidemiological diagnosis of asthma. Eur Respir J. 1998;11(3):599–605.
67. Menezes AMB, Perez-Padilla R, Jardim JRB, Muiño A, Lopez MV, Valdivia G, et al. Chronic obstructive pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. Lancet. 2005;366(9500):1875–81.
68. Hanania NA, Noonan M, Corren J, Korenblat P, Zheng Y, Fischer SK, Cheu M, Putnam WS, Murray E, Scheerens H, et al. Lebrikizumab in moderate-to-severe asthma: pooled data from two randomised placebo-controlled studies. Thorax. 2015;70(8):748–56.
69. Corren J, Lemanske RF, Hanania NA, Korenblat PE, Parsey MV, Arron JR, Harris JM, Scheerens H, Wu LC, Su Z, et al. Lebrikizumab treatment in adults with asthma. N Engl J Med. 2011;365(12):1088–98.
70. Dweik RA, Boggs PB, Erzurum SC, Irvin CG, Leigh MW, Lundberg JO, Olin AC, Plummer AL, Taylor DR; American Thoracic Society Committee on Interpretation of Exhaled Nitric Oxide Levels (FENO) for Clinical Applications. An official ATS clinical practice guideline: interpretation of exhaled nitric oxide levels (FENO) for clinical applications. Am J Respir Crit Care Med. 2011;184(5):602–15.
71. Global Strategy for Asthma Management and Prevention 2016. http://ginasthma.org/2016-gina-report-global-strategy-for-asthma-management-and-prevention/. Accessed 12 May 2017.

Appendix III

Study II

Amaral, R., Pereira, A. M., Jacinto, T., Malinovschi, A., Janson, C., Alving, K., & Fonseca, J.A. (2019). Comparison of hypothesis- and data-driven asthma phenotypes in NHANES 2007-2012: the importance of comprehensive data availability. *Clinical and Translational Allergy, 9*:17. doi.org/10.1186/s13601-019-0258-7. eCollection 2019.

Amaral et al. Clin Transl Allergy (2019) 9:17
https://doi.org/10.1186/s13601-019-0258-7

Clinical and Translational Allergy

RESEARCH Open Access

Comparison of hypothesis- and data-driven asthma phenotypes in NHANES 2007–2012: the importance of comprehensive data availability

Rita Amaral[1,2*], Ana M. Pereira[1,3], Tiago Jacinto[1,2], Andrei Malinovschi[4], Christer Janson[5], Kjell Alving[6] and João A. Fonseca[1,3,7]

Abstract

Background: Half of the adults with current asthma among the US National Health and Nutrition Examination Survey (NHANES) participants could be classified in more than one hypothesis-driven phenotype. A data-driven approach applied to the same subjects may allow a more useful classification compared to the hypothesis-driven one.

Aim: To compare previously defined hypothesis-driven with newly derived data-driven asthma phenotypes, identified by latent class analysis (LCA), in adults with current asthma from NHANES 2007–2012.

Methods: Adults (≥ 18 years) with current asthma from the NHANES were included (n = 1059). LCA included variables commonly used to subdivide asthma. LCA models were derived independently according to age groups: < 40 and ≥ 40 years old.

Results: Two data-driven phenotypes were identified among adults with current asthma, for both age groups. The proportions of the hypothesis-driven phenotypes were similar among the two data-driven phenotypes (p > 0.05). Class A < 40 years (*n* = 285; 75%) and Class A ≥ 40 years (n = 462; 73%), respectively, were characterized by a predominance of highly symptomatic asthma subjects with poor lung function, compared to Class B < 40 years (n = 94; 25%) and Class B ≥ 40 years (n = 170; 27%). Inflammatory biomarkers, smoking status, presence of obesity and hay fever did not markedly differ between the phenotypes.

Conclusion: Both data- and hypothesis-driven approaches using clinical and physiological variables commonly used to characterize asthma are suboptimal to identify asthma phenotypes among adults from the general population. Further studies based on more comprehensive disease features are required to identify asthma phenotypes in population-based studies.

Keywords: Asthma, Phenotypes, Population-based study, Unsupervised analysis

*Correspondence: rita.s.amaral@gmail.com
[1] CINTESIS – Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Edifício Nascente, Piso 2, Rua Dr. Plácido da Costa, s/n, 4200-450 Porto, Portugal
Full list of author information is available at the end of the article

Introduction

Airways diseases, such as asthma and chronic obstructive pulmonary disease (COPD), comprise a heterogeneous set of subtypes with different underlying pathophysiological mechanisms [1–3]. Both hypothesis-driven and data-driven methods can be used to classify patients into sub-groups of airways diseases [4–6].

The hypothesis-driven approach classifies airways diseases based on pre-defined criteria following immunopathology concepts and asthma literature, while in data-driven methods no prior disease classification is required [7, 8]. Data-driven approaches have provided insights into "novel" phenotypes of complex disease pathogenesis, suggesting disease stratification depending on the individual pathophysiologic characteristics [8–11].

Most studies on asthma phenotyping using data-driven methods emphasize patients with moderate to severe asthma and/or clinically-based settings [12–15]. Therefore, the generalization to the general asthma population may be limited.

Different types of data-driven methods have been widely used in airway diseases, such as hierarchical [12], partitioning [14], and latent class analysis (LCA) [10]. Notably, LCA appeared to account better for the heterogeneity of airways symptoms, compared to other commonly used data-driven approaches (e.g. partitioning around medoids) [16]. Moreover, the application of the latent class assignments developed from a national data source has previously demonstrated higher degrees of generalizability [17].

Recently, we reported a significant overlap between five distinct hypothesis-driven asthma phenotypes in adults from the general population included in the US National Health and Nutrition Examination Survey (NHANES) [18]. We have emphasized that a combination of clinical information and biomarkers, using a more comprehensive data analysis approach, such as data-driven methods, could provide a better taxonomy of non-severe asthma.

In this study, we aimed to compare previously defined hypothesis-driven asthma phenotypes [18] with data-driven asthma phenotypes derived by applying LCA to a sample of adults representative of the US general population.

Methods

Study setting and participants

We have included subjects that participated in the NHANES study, a nationally representative survey of the civilian, non-institutionalized US population performed with the aim of gathering data regarding health and nutritional status. Protocols were approved by the National Center for Health Statistics Research Ethics Review Board and all participants gave written informed consent. Detailed information can be found in the NHANES documentation (www.cdc.gov/nchs/nhanes.htm).

Data from three NHANES surveys was used (n = 30,442). We included adults (≥ 18 years old) with current asthma (n = 1059), defined by a positive answer to the questions [18]: "Has a doctor ever told you that you have asthma?" together with "Do you still have asthma?", and either "wheezing/whistling in the chest in the past 12 months" or "asthma attack in the past 12 months."

Variables

Anthropometric and demographic characteristics, such as age, gender, body mass index (BMI), and smoking status were analysed, as well as blood eosinophils (B-Eos) count, fraction of exhaled nitric oxide (FeNO) and spirometric parameters. FeNO and spirometry were performed following ATS/ERS recommendations [19, 20]. Basal predicted values of forced expiratory volume during the first second (FEV_1) and forced vital capacity (FVC) were calculated [21, 22] and abnormal values were defined as being below the lower limit of normal (LLN) [23].

Hypothesis-driven asthma phenotypes

The analysis based on the report of smoking status, presence of obesity and inflammatory markers enabled the definition of five asthma phenotypes [18]: B-Eos-high asthma phenotype, if B-Eos ≥ 300/mm^3; FeNO-high asthma, if FeNO ≥ 35 ppb; B-Eos&FeNO-low asthma, if B-Eos < 150/mm^3 and FeNO < 20 ppb; asthma with obesity (AwObesity), if BMI ≥ 30 kg/m^2; and asthma with concurrent COPD (AwCOPD), if subjects had self-reported chronic bronchitis/emphysema with age of diagnosis ≥ 40 years and being either a current or an ex-smoker (ever smoked). Subjects were considered as "non-classified" if they did not meet the criteria for any of the defined asthma phenotypes. Additionally, to account for individuals with probable co-existence of asthma and COPD and minimize age as a confounding variable, we conducted the analysis considering two age groups: < 40 and ≥ 40 years old [18].

Data-driven asthma phenotypes

LCA was used to identify asthma phenotypes in an unsupervised manner (data-driven approach). Two models for "current asthma" were developed (Additional file 1: Table S1): Model 1 was based on the 4 variables previously used to define the hypothesis-driven asthma phenotypes (BMI ≥ 30 kg/m^2, ever-smoking status, FeNO ≥ 35 ppb, B-Eos ≥ 300/mm^3) [18]; and in Model 2, we have added to the former 4 variables, sex, early asthma onset (< 16 years old), wheezing-related questions

(presence/absence of at least one wheezing attack, wheezing with exercise, sleep disturbance by wheezing, limit activity by wheezing, absenteeism by wheezing), asthma-related emergency department (ED) visit in the previous 12 months, $FEV_1/FVC < LLN$, $FEV_1 < LLN$, and self-reported hay fever.

Additionally, to explore the results in different "asthma populations", we've developed two other models using similar variables. For the "ever asthma" subgroup (model 3) we included subjects with a positive answer to "Has a doctor ever told you that you have asthma?" (n = 2611); and for the "difficult asthma" (model 4) we included subjects with poor asthma-related outcomes, defined as current asthma plus, at least, one of the following: asthma-related ED visit, $FEV_1 < LLN$, or oral corticosteroids use in the past 30 days (n = 673) (Additional file 1: Table S1).

Latent class models were derived independently for each age group, using the same variables, and a secondary analysis without stratifying by age was done on the three asthma subgroups. The most appropriate number of clusters was determined by examining commonly used criteria [24]. Further methodological details are found in the Additional file 1.

Statistical analysis

All analyses considered the complex multistage sampling and 6-year sampling weights provided by the NHANES documentation [25]. LCA was performed with MPlus (version 6.12), that considered the complex survey design of NHANES when performing LCA-modelling. All other analysis was performed in Stata/IC 15.1 (Stata Corp, College Station, TX, USA). A p-value < 0.05 was considered statistically significant.

Results

We included 1059 adults with current asthma. The weighted proportions of the previously defined hypothesis-driven asthma phenotypes, according to age groups (< 40 and ≥ 40 years old) were, respectively: 42% and 53% with AwObesity; 34% and 37% with B-Eos-high asthma; 26% and 21% for B-Eos&FeNO-low; 18% and 19% with FeNO-high asthma; and 19% AwCOPD, in the older group [18]. In addition, 17% and 12% of the individuals in the < 40 and ≥ 40 years old groups, respectively, were categorized as "non-classified".

In Model 1, LCA was not able to differentiate any asthma subgroup among subjects with current asthma (Additional file 1: Table S1). On the other hand, by adding more asthma-related variables (Model 2), LCA identified a two-class model as the best solution for both age groups (Table 1, Additional file 1: Table S1). Classes A < 40 years (n = 290; 75%) and A ≥ 40 years (n = 494; 73%) had marked predominance of highly symptomatic asthma subjects, with poorer lung function, compared to classes B < 40 years (n = 96; 25%) and B ≥ 40 years (n = 179; 27%), respectively (Table 1). Regarding inflammatory markers, the proportion of patients with high levels of B-Eos and FeNO was not significantly different between classes, both in the younger group ($p = 0.99$ and $p = 0.82$, respectively) and in the older group ($p = 0.57$ and $p = 0.53$).

Figure 1 shows that the distribution of the hypothesis-driven phenotypes is similar ($p > 0.05$) in both classes identified by LCA regardless age group.

Additionally, LCA identified 2 classes on the models for "ever-asthma" and "current asthma" without stratifying by age, but not for the difficult-asthma sub analysis where no subgroup was identified (Additional file 1: Table S1).

Discussion

This was the first study comparing previously defined hypothesis-driven asthma phenotypes with data-driven ones in a sample representative of the US general population. The proportions of the hypothesis-driven phenotypes were similar between the two data-driven phenotypes obtained by LCA using clinical and physiological variables commonly used to characterize asthma.

Previous studies using data-driven approaches contributed to the definition of clusters/phenotypes based on similarities in clinical and inflammatory biomarkers [9, 12–14]. However, these approaches have been scarcely applied to adults with asthma from population-based studies. The studies from Siroux et al. [26] and Mākikyrō et al. [27] provided further evidence for identifying subgroups of asthma based on clinical markers and questionnaire data commonly available in primary health care or large epidemiological studies and found a larger range of asthma phenotypes.

Our study showed that performing LCA with the variables used to define some of the most common hypothesis-driven asthma phenotypes, could not identify subgroups within adults with current asthma from the general population. By including additional clinical and physiological variables commonly used to classify asthma, LCA identified two data-driven phenotypes in the same subjects. Overall, these phenotypes only differed in symptom frequency and lung function parameters. Inflammatory biomarkers, presence of obesity, smoking status, age of asthma onset and self-reported hay fever were not different between classes.

Moreover, using a less stringent asthma definition (ever asthma) and in subjects with poor clinical outcomes (difficult asthma), these variables were also suboptimal to differentiate asthma subgroups.

Table 1 Proportions of each variable according to the LCA-classes identified in Model 2 (subjects with current asthma, n = 1059)

Class A<40 years n=280 (75%)	Class B<40 years n=98 (25%)	Variables	Class A≥40 years n=494 (73%)	Class B≥40 years n=179 (27%)
99%	0%	Limit activity by wheezing	88%	2%
76%	0%	Wheezing with exercise	88%	8%
84%	0%	Sleep disturbance by wheezing	63%	0%
88%	34%	Wheezing attack	99%	45%
29%	10%	FEV_1/FVC <LLN	34%	16%
19%	0%	Absenteeism by wheezing	16%	0%
28%	14%	ED by asthma	20%	11%
41%	59%	Female	66%	55%
20%	11%	FEV_1 <LLN	35%	19%
16%	25%	FeNO ≥35 ppb	21%	15%
34%	36%	B-Eos ≥300/mm^3	40%	30%
47%	41%	Ever smoked	61%	58%
26%	20%	Hay fever	44%	46%
43%	41%	BMI ≥30kg/m^2	55%	52%
51%	50%	Early asthma onset	25%	22%

FEV_1 forced expiratory volume in the first second, *FVC* forced vital capacity, *LLN* lower limit of normality, *ED* emergency department, *FeNO* fractional exhaled nitric oxide, *B-Eos* blood eosinophils count, *BMI* body mass index

Variables are ordered by the highest mean difference between the 2 classes of each age group and each coloured box represents the prevalence of the variables within the class, ranging from 0% (light yellow) to 100% (red)

In contrast to studies with severe asthma patients, our results suggest that, for the general asthma population, the clinical and physiological variables available to classify asthma and commonly used predefined cut-offs seem to be insufficient to identify specific phenotypes. The inclusion in data-driven models of additional easily measurable biomarkers that have already been shown to be helpful in discriminating asthma phenotypes in this population (e.g. serum IgE and/or periostin) [28, 29], combined with comprehensive clinical, physiologic, and/or disease features, might result in the identification of more precise phenotypes. Also, the identification of new, more accurate biomarkers could also improve phenotyping [30]. Furthermore, the use of fixed cut-offs values, although common and more intuitive for daily clinical practice, may potentially miss more complex, and yet unidentified phenotypes. The use of absolute values (as seen in other studies [13, 31, 32]), or appropriate reference equations for predicted values [33, 34] could be more adequate.

Similarly, research efforts are being made to integrate clinical characteristics with available biomarkers to identify data-driven asthma phenotypes in children [35, 36]. However, the obtained phenotypes vary on key features that are more pronounced during childhood, including natural history of wheeze over time [37], suggesting that further work is required to compare data- and hypothesis-driven approaches to identify asthma phenotypes in children.

Limitations inherent to a survey study design must be acknowledged and the self-reported variables may lead to misclassifications and information biases; to account for these biases, we used previously validated definitions [38, 39]. Also, despite including the most commonly used variables for respiratory disease assessment available in the NHANES study, when using the less stringent asthma definition, the differentiation of asthma subgroups was not improved in this population. However, to reduce the risk of poor LCA-class differentiation, we did not include any of the variables used in the asthma groups definition into the LCA models. Finally, LCA modelling should comprehend all the domains relevant to the understanding of the disease to classify observations into discrete and mutually exclusive classes [40], suggesting that the use of predefined cut-offs and the lack of data regarding, for example, objective assessment of atopy, nasal and ocular symptoms (which have proved to be useful in the stratification of allergic respiratory diseases [10, 41]), may have limited the ability to differentiate specific asthma phenotypes using unsupervised analysis.

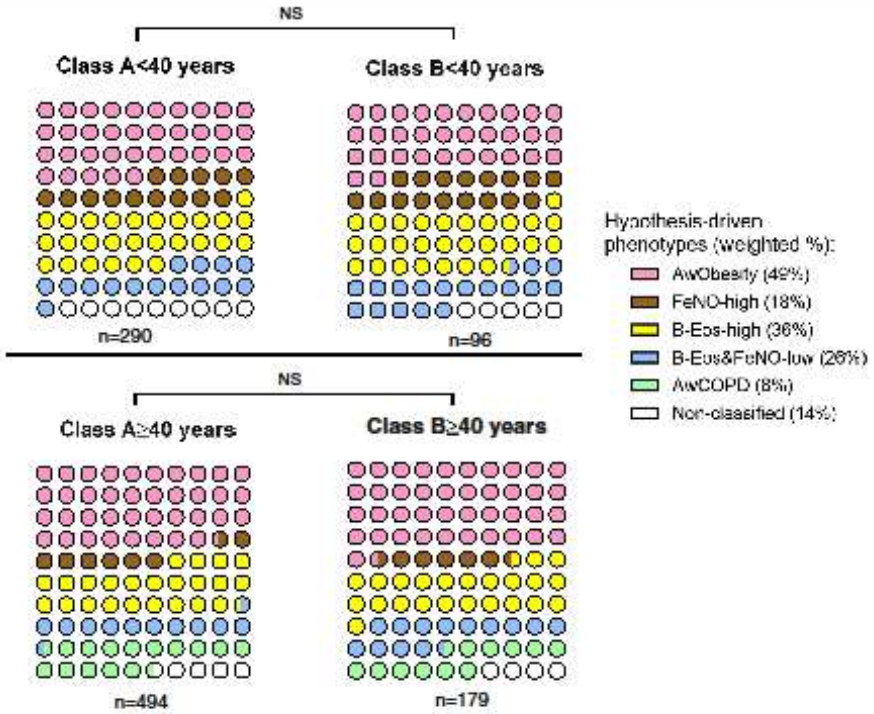


Fig. 1 Distribution of the hypothesis-driven asthma phenotypes according to the data-driven classes identified in Model 2. Both Class A < 40 and Class A ≥ 40 are the phenotypes with more asthma-related symptoms and low lung function. No significant differences ($p > 0.05$) were observed between the proportions of the hypothesis-driven within the data-driven phenotypes. *NS* non-significant

In conclusion, this brief communication extends our previous work on the need for a broader data analysis combining different asthma-related domains for differentiating phenotypes in the general asthma population [18]. The clinical and physiological variables commonly used to subdivide asthma seem to be insufficient to differentiate specific asthma phenotypes among adults from the general population, irrespective of using data-driven or hypothesis-driven approaches. Further studies based on more comprehensive disease features are required to identify asthma phenotypes with the potential to be useful for clinicians and for population-based research.

Additional file

Additional file 1. Supplementary methods.

Abbreviations

ATS/ERS: American Thoracic Society/European Respiratory Society; AwCOPD: asthma with concurrent COPD; AwObesity: asthma with obesity; B-Eos: blood eosinophils; BMI: body mass index; COPD: chronic obstructive pulmonary disease; ED: emergency department; FeNO: fraction of exhaled nitric oxide; FEV_1: forced expiratory volume during the first second; FVC: forced vital capacity; LCA: latent class analysis; LLN: lower limit of normal; NHANES: National Health and Nutrition Examination Survey; US: United States.

Authors' contributions

RA, AMP, JAF, contributed to study conception and design, analysis and interpretation of data, writing and revising the article. TJ, AM, CJ and KA contributed to data interpretation, writing and revising the article. All authors read and approved the final manuscript.

Author details

[1] CINTESIS – Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Edifício Nascente, Piso 2, Rua Dr. Plácido da Costa, s/n, 4200-450 Porto, Portugal. [2] Department of Cardiovascular and Respiratory Sciences, Porto Health School, Porto, Portugal. [3] Department of Allergy, Instituto & Hospital CUF, Porto, Portugal. [4] Department of Medical Sciences, Clinical Physiology, Uppsala University, Uppsala, Sweden. [5] Department of Medical Sciences, Respiratory Medicine and Allergology, Uppsala University, Uppsala, Sweden. [6] Department of Women's and Children's Health, Paediatric Research, Uppsala University, Uppsala, Sweden. [7] MEDCIDS

- Department of Community Medicine, Information, and Health Sciences, Faculty of Medicine, University of Porto, Porto, Portugal.

Acknowledgements
This article was supported by FEDER through the operation POCI-01-0145-FEDER-007746 funded by the Programa Operacional Competitividade e Internacionalização – COMPETE2020 and by National Funds through FCT – Fundação para a Ciência e a Tecnologia within CINTESIS, R&D Unit (reference UID/IC/4255/2013).

Competing interests
The authors declare that they have no competing interests.

Availability of data
Data and respective datasets are displayed at the NHANES website: https://www.cdc.gov/nchs/nhanes/Index.htm.

Consent of publication
Not applicable.

Ethics approval and consent to participate
The NHANES survey operates under the approval of the National Center for Health Statistics Research Ethics Review Board (Protocols #2005-06, and #2011-17), available in www.cdc.gov/nchs/nhanes/irba98.htm. All the NHANES data meet the conditions described in Research Using Publicly Available Datasets (Secondary Analysis) - Policy #39 - for use without application to Institutional Review Board. All study participants provided written informed consent.

Funding
RA is supported by a Ph.D. grant (grant no. PD/BD/113659/2015), financed by the *Fundação para a Ciência e Tecnologia, I.P.*, PhD program (reference no. PD/0003/2013: Doctoral Program in Clinical and Health Services Research).

Publisher's Note

Received: 4 January 2019 Accepted: 1 March 2019
Published online: 13 March 2019

References

1. Pavord ID, Beasley R, Agusti A, Anderson GP, Bel E, Brusselle G, et al. After asthma: redefining airways diseases. Lancet. 2018;391(10118):350–400.
2. Pavord ID, Shaw DE, Gibson PG, Taylor DR. Inflammometry to assess airway diseases. Lancet. 2008;372(9643):1017–9.
3. Wurst KE, Kelly-Reif K, Bushnell GA, Pascoe S, Barnes N. Understanding asthma-chronic obstructive pulmonary disease overlap syndrome. Respir Med. 2016;110:1–11.
4. Wenzel S. Asthma: defining of the persistent adult phenotypes. Lancet. 2006;368:804–13.
5. Bousquet J, Anto JM, Stark PJ, Adcock IM, Chung KF, Roca J, et al. Systems medicine and integrated care to combat chronic noncommunicable diseases. Genome Med. 2011;3(7):43.
6. Prosperi MCF, Sahiner UM, Belgrave D, Sackesen C, Buchan IE, Simpson A, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. Am J Respir Crit Care Med. 2013;188(11):1303–12.
7. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Waltham: Morgan Kaufmann Publishers; 2012.
8. Yii ACA, Tay T-R, Choo XN, Koh MSY, Tee AKH, Wang D-Y. Precision medicine in united airways disease: a "treatable traits" approach. Allergy. 2018;73(10):1964–78.
9. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med. 2008;178(3):218–24.
10. Amaral R, Bousquet J, Pereira AM, Araújo LM, Sá-Sousa A, Jacinto T, et al. Disentangling the heterogeneity of allergic respiratory diseases by latent class analysis reveals novel phenotypes. Allergy. 2018. https://doi.org/10.1111/all.13670.
11. Anto JM, Bousquet J, Akdis M, Auffray C, Keil T, Momas I, et al. Mechanisms of the Development of Allergy (MeDALL): Introducing novel concepts in allergy phenotypes. J Allergy Clin Immunol. 2017;139(2):388–99.
12. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med. 2010;181(4):315–23.
13. Wu W, Bleecker E, Moore W, Busse WW, Castro M, Chung KF, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. J Allergy Clin Immunol. 2014;133(5):1280–8.
14. Lefaudeux D, De Meulder B, Loza MJ, Peffer N, Rowe A, Baribaud F, et al. U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. J Allergy Clin Immunol. 2017;139(6):1797–807.
15. Amelink M, de Nijs SB, de Groot JC, van Tilburg PMB, van Spiegel PI, Krouwels FH, et al. Three phenotypes of adult-onset asthma. Allergy. 2013;68(5):674–80.
16. Amaral R, Jacinto T, Pereira A, Almeida R, Fonseca J. A comparison of unsupervised methods based on dichotomous data to identify clusters of airways symptoms: latent class analysis and partitioning around medoids methods. Eur Respir J. 2018; https://doi.org/10.1183/13993003.congress-2018.PA4429.
17. Evenson KR, Wen F, Howard AG, Herring AH. Applying latent class assignments for accelerometry data to external populations: data from the National Health and Nutrition Examination Survey 2003–2006. Data Br. 2016;9:926–30.
18. Amaral R, Fonseca JA, Jacinto T, Pereira AM, Malinovschi A, Janson C, et al. Having concomitant asthma phenotypes is common and independently relates to poor lung function in NHANES 2007–2012. Clin Transl Allergy. 2018;8(1):13.
19. Silkoff PE. ATS/ERS recommendations for standardized procedures for the online and offline measurement of exhaled lower respiratory nitric oxide and nasal nitric oxide, 2005. Am J Respir Crit Care Med. 2005;171(8):912–30.
20. Miller MR. Standardisation of spirometry. Eur Respir J. 2005;26(2):319–38.
21. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med. 1999;159(1):179–87.
22. Hankinson JL, Kawut SM, Shahar E, Smith LJ, Stukovsky KH, Barr RG. Performance of American thoracic society-recommended spirometry reference values in a multiethnic sample of adults. Chest. 2010;137(1):138–45.
23. Stanojevic S, Wade A, Stocks J, Hankinson J, Coates AL, Pan H, et al. Reference ranges for spirometry across all ages. Am J Respir Crit Care Med. 2008;177(3):253–60.
24. Muthén LK, Muthén BO. Mplus user's guide. 7th ed. Los Angeles: Muthén & Muthén; 2012.
25. Specifying weightning parameters. https://www.cdc.gov/nchs/tutorials/nhanes/SurveyDesign/Weighting/intro.htm. Accessed 9 Dec 2018.
26. Siroux V, Basagaña X, Boudier A, Pin I, Garcia-Aymerich J, Vesin A, et al. Identifying adult asthma phenotypes using a clustering approach. Eur Respir J. 2011;38(2):310–7.
27. Mäkikyrö EMS, Jaakkola MS, Jaakkola JJK. Subtypes of asthma based on asthma control and severity: a latent class analysis. Respir Res. 2017;18(1):24.
28. Patelis A, Gunnbjörnsdottir M, Malinovschi A, Matsson P, Önell A, Högman M, et al. Population-based study of multiplexed IgE sensitization in relation to asthma, exhaled nitric oxide, and bronchial responsiveness. J Allergy Clin Immunol. 2012;130(2):397–402.e2.
29. James A, Janson C, Malinovschi A, Holweg C, Alving K, Ono J, et al. Serum periostin relates to type-2 inflammation and lung function in asthma: data from the large population-based cohort Swedish GA(2)LEN. Allergy. 2017;72(11):1753–60.
30. Carr TF, Kraft M. Use of biomarkers to identify phenotypes and endotypes of severe asthma. Ann Allergy Asthma Immunol. 2018;121(4):414–20.
31. Hsiao H-P, Lin M-C, Wu C-C, Wang C-C, Wang T-N. Sex-specific asthma phenotypes, inflammatory patterns, and asthma control in a cluster analysis. J Allergy Clin Immunol Pract. 2019;7(2):556–567.e15.
32. Sendín-Hernández MP, Ávila-Zarza C, Sanz C, García-Sánchez A, Marcos-Vadillo E, Muñoz-Bellido FJ, et al. Cluster analysis identifies 3 phenotypes within allergic asthma. J Allergy Clin Immunol Pract. 2018;6(3):955–961.e1.

33. Quanjer P, Stanojevic S. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. Eur Respir J. 2012;40:1324–43.
34. Jacinto T, Amaral R, Malinovschi A, Janson C, Fonseca J, Alving K. Exhaled NO reference limits in a large population-based sample using the Lambda-Mu-Sigma method. J Appl Physiol. 2018;125(5):1620–6.
35. Dapnar M, Fuchs O, Genuneit J, Karvonen AM, Hyvärinen A, Kaulek V, et al. Clinical and epidemiologic phenotypes of childhood asthma. Am J Respir Crit Care Med. 2014;189(2):129–38.
36. Collins SA, Pike KC, Inskip HM, Godfrey KM, Roberts G, Holloway JW, et al. Validation of novel wheeze phenotypes using longitudinal airway function and atopic sensitization data in the first 6 years of life: evidence from the Southampton Women's survey. Pediatr Pulmonol. 2013;48(7):683–92.
37. Henderson J, Granell R, Heron J, Sherriff A, Simpson A, Woodcock A, et al. Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood. Thorax. 2008;63(11):974–80.
38. Sá-Sousa A, Jacinto T, Azevedo LF, Morais-Almeida M, Robalo-Cordeiro C, Bugalho-Almeida A, et al. Operational definitions of asthma in recent epidemiological studies are inconsistent. Clin Transl Allergy. 2014;4:24.
39. Halldin CN, Doney BC, Hnizdo E. Changes in prevalence of chronic obstructive pulmonary disease and asthma in the US population and associated risk factors. Chron Respir Dis. 2015;12(1):47–60.
40. Wang J, Wang X. Structural equation modeling: applications using Mplus. West Sussex: Wiley; 2012.
41. Bousquet J, Devillier P, Anto JM, Bewick M, Haahtela T, Arnavielhe S, et al. Daily allergic multimorbidity in rhinitis using mobile technology: a novel concept of the MASK study. Allergy. 2018;73(8):1622–31.

Appendix IV

Study III

Amaral, R., Bousquet, J., Pereira, A.M., Araújo, L., Sá-Sousa, A., Jacinto, T., Almeida, R., Delgado, L. & Fonseca, J.A. (2019). Disentangling the heterogeneity of allergic respiratory diseases by latent class analysis reveals novel phenotypes". *Allergy, 74*(4), 698-708. doi: 10.1111/all.1367 0. Epub 2018 Dec 5.

Received: 30 July 2018 | Revised: 23 October 2018 | Accepted: 6 November 2018

DOI: 10.1111/all.13670

ORIGINAL ARTICLE

Asthma and Lower Airway Disease

WILEY Allergy

Disentangling the heterogeneity of allergic respiratory diseases by latent class analysis reveals novel phenotypes

Rita Amaral[1,2] | Jean Bousquet[3,4,5,6,7] | Ana M. Pereira[1,8] | Luís M. Araújo[8,9] | Ana Sá-Sousa[1] | Tiago Jacinto[1,2] | Rute Almeida[1] | Luís Delgado[8,9] | Joao A. Fonseca[1,8,10]

[1]CINTESIS - Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Porto, Portugal

[2]Department of Cardiovascular and Respiratory Sciences, Porto Health School, Porto, Portugal

[3]MACVIA-France, Contre les MAladies Chroniques pour un Vieillissement Actif en France European Innovation Partnership on Active and Healthy Ageing Reference Site, Montpellier, France

[4]INSERM U 1168, VIMA: Ageing and Chronic Diseases Epidemiological and Public Health Approaches, Villejuif, France

[5]UMR-S 1168, Université Versailles St-Quentin-en-Yvelines, Montigny le Bretonneux, France

[6]Euforea, Brussels, Belgium

[7]Euforea, Charité, Germany

[8]Allergy Unit, Instituto & Hospital CUF, Porto, Portugal

[9]Basic & Clinical Immunology, Pathology Department, Faculdade Medicina da, Universidade do Porto, Porto, Portugal

[10]MEDCIDS, Department of Community Medicine, Information, and Health Sciences, Faculty of Medicine, University of Porto, Porto, Portugal

Correspondence
Rita Amaral, CINTESIS - Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Porto, Portugal.
Email: rita.s.amaral@gmail.com

Funding Information
Fundação para a Ciência e a Tecnologia, Grant/Award Number: PD/BD/113659/2015, PTDC/SAU-SAP/119192/2010

Abstract

Background: Refined phenotyping of allergic diseases may unravel novel phenotypes. Conjunctivitis as an independent disorder has never been approached.

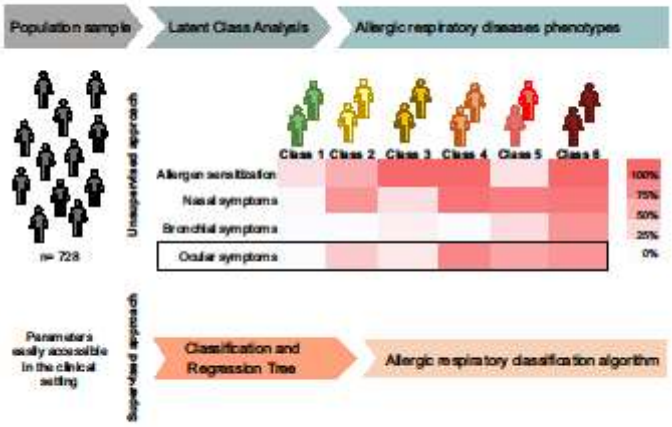
Aim: To identify distinct classes of allergic respiratory diseases using latent class analysis (LCA) and distinguish each class using classification and regression tree (CART) analysis.

Methods: Seven hundred and twenty-eight adults from the Portuguese general population study ICAR had a structured medical interview combined with blood collection, skin prick tests, spirometry with bronchodilation, and exhaled nitric oxide. LCA was applied to 19 variables. The CART algorithm selected the most likely variables distinguishing LCA-classes.

Results: A six-class model was obtained. Class 1 (25%): nonallergic participants without bronchial or ocular symptoms. Classes 2 (22%) and 3 (11%): nasal and ocular (low levels) symptoms without nasal impairment, monosensitized (Class 2) or polysensitized (Class 3). Class 4 (13%): polysensitized participants with high levels of nasal and ocular symptoms, and nasal impairment. Classes 5 (16%) and 6 (14%): high level of nasal, bronchial and ocular symptoms with nasal impairment (non-allergic or polysensitized, respectively). Participants in classes 5 and 6 had more bronchial exacerbations and unscheduled medical visits ($P < 0.001$). Ocular symptoms were significantly higher in classes with nasal impairment, compared to those without impairment ($P < 0.001$) or no nasal symptom ($P < 0.001$). CART highlighted ocular symptoms as the most relevant variable in distinguishing LCA-classes.

Conclusion: Novel severe phenotypes of participants with co-occurrence of ocular, nasal and bronchial symptoms, and exacerbation-prone were identified. The tree algorithm showed the importance of the ocular symptoms in the expression of allergic diseases phenotypes.

KEYWORDS
airways symptoms, allergic sensitization, cluster analysis, ocular symptoms, phenotype



GRAPHICAL ABSTRACT

Novel phenotypes of allergic respiratory diseases were identified by latent class analysis, in adults from a general population sample. For the first time, ocular symptoms were used independently and ranked by classification and regression tree analysis as the most relevant variable to separate those phenotypes.

1 | INTRODUCTION

Allergic respiratory diseases (ARD) such as asthma, allergic rhinitis, and conjunctivitis, often co-exist in the same patient.[1] Allergic multimorbidity is independent of Immunoglobulin E (IgE) sensitization,[2] but the multimorbid polysensitized phenotype is among the most severe ones.[3-5]

Multimorbidity is complex and may be difficult to manage in clinical practice.[6] Disease stratification using classical diagnostic methods may not be sufficient for a more effective and personalized patient care. Identification of "treatable traits" by data-driven approaches (unsupervised methods) may be a promising approach[7,8] using machine learning tools.[9,10] Previous studies used data-clustering (unsupervised) methods, such as latent class analysis (LCA), to identify separately phenotypes of asthma,[11,12] rhinitis,[13,14] and other allergic diseases,[15,16] but nasal and ocular symptoms were considered together.[16]

The LCA seeks to identify homogeneous groups in a heterogeneous descriptive, a theoretical, and non-inferential manner without providing information on the comparative relevance of variables or on the classification decision rule.[17] The classification and regression tree (CART) analysis can develop reliable clinical decision rules to be used in the development of new classification of patients into categories,[18,19] because of their easily interpretive nature and ability to handle missing data.[19] The development of a classification tree to classify participants in the ARD phenotypes, using data readily accessible in most outpatient settings, is expected to improve patient classification and management.

1.1 | Aim

The present cross-sectional study aimed to identify distinct classes of allergic respiratory diseases, in adults from a general population sample, using latent class analysis and to explore the most relevant clinical variables that could be used to distinguish each class, using classification and regression tree analysis. Additionally, we included ocular symptoms that are usually not considered independently in allergic multimorbidity but were supported by a study using mobile technology[20] and suggested by MeDALL.[16]

2 | METHODS

2.1 | Sample and study design

This is a secondary analysis of a national and cross-sectional study conducted in the Portuguese general population, ICAR (Control and Burden of Asthma and Rhinitis) study.[21] It comprised anthropometric measurements, lung function, inflammation and allergy tests, a structured clinical assessment and standardized questionnaires. Detailed information is provided in the Online Supplement. ICAR study enrolled 858 participants, from 3 to 89 years old, with and without self-reported asthma and/or rhinitis (Figure S1) and, for this analysis, we considered all adults (≥18 years old, n = 728) (Table S1).

Ethical approval was obtained from the Hospital Ethics Committee (Comissão de Ética do Hospital São João, E.P.E) and national data protection committee (n.12372/2011). All participants gave written informed consent. The paper follows the STROBE guidelines for reporting observational studies.[22]

2.2 | Variables

Demographic characteristics, such as age, gender, body mass index (BMI), and smoking status were analyzed. Data regarding nasal, bronchial

and ocular symptoms, bronchial exacerbation and unscheduled medical visit, in the last 12 months, were collected by clinical interview. Age at onset of the bronchial symptoms was self-reported. Further details on variable definition are described in Online Supplement.

Nasal severity score was adapted from the ARIA severity score,[20] and it was calculated using questions regarding impact of nasal symptoms on daily activities, work, and sleep, ascribed with the score 1 if "Yes" and 0 if "No." The nasal severity score was then categorized as "no/mild impairment" (ranging 0-2) and "severe impairment" (score = 3).

Additionally, a diagnosis of rhinitis, asthma, and other allergic diseases was established by an allergy specialist and the use of asthma/allergy medication in the last 12 months was also analyzed.

2.3 | Measurements

Allergic sensitization (AS) was assessed by skin-prick tests (SPT), following the GA2LEN recommendations.[23] The standardized allergen panel included 28 allergens (Stallergenes Greer®, France),[24] categorized into six groups (mites, dog and cat epitheliums, tree, grass and weed pollens mixtures, and molds). Monosensitization and polysensitization were defined, respectively, as sensitization to only one and to two or more groups of allergens. Furthermore, sensitization data of 25 participants with current antihistamine medication were considered as missing.

Total immunoglobulin E (IgE), serum eosinophilic cationic protein (S-ECP) and blood eosinophilic count (B-Eos) were obtained from blood sampling. Phadiatop® (Thermo Fisher Scientific, Uppsala, Sweden) was used as a screening test; if ≥0.35kU/l the sample was considered as Phadiatop-positive, and additional determinations were performed to assess individual allergen-specific IgE (sIgE) antibody concentrations.

Predicted values of basal forced expiratory volume in one second (FEV_1) and bronchodilator response were obtained by spirometry procedure, following the ATS/ERS recommendations:[25,26] FEV_1 were considered abnormal if less than the lower limit of normal (LLN).[27] Fractional exhaled Nitric Oxide (FeNO) measurements were performed using NIOX MINO® (Aerocrine, Solna, Sweden), following the ATS/ERS criteria.[28]

2.4 | Biases

To reduce the risk of bias, several quality assurance measures were followed: research assistants performing the evaluations were blinded to the subject classification at screening; data validity was periodically verified soon after being collected and custom statistic algorithms were used to detect extreme, illogical and missing value and amendments to the protocol were done if necessary.

2.5 | Statistical analyses

Categorical variables are presented as absolute frequencies and proportions. Continuous variables were presented according to their distributions. The socio-demographic and clinical variables of the classes derived from the LCA were described and compared using chi-squared test, one-way analysis of variance and Kruskal-Wallis test with Bonferroni correction.

Mplus 6.12 (Los Angeles, CA:Muthén & Muthén) was used to conduct LCA analysis and R 3.3.3 (https://www.r-project.org/) to establish the classification model and build the respective decision tree, using the "rpart" and "DMwR" packages, respectively. All other analyses were performed using SPSS Statistics 25.0 (Armonk, NY:IBM Corp) and *P*-values < 0.05 were defined as statistically significant.

2.6 | Unsupervised analysis

LCA was applied to identify underlying unobserved (latent), mutually exclusive subgroups (classes) based on categorical manifest variables without the need for historical or *a priori* assumptions.[29] Nineteen dichotomic variables (defined as "Yes"/"No") were chosen regarding nasal, ocular and bronchial symptoms and the 6 groups of AS.

2.7 | Supervised analysis

CART analysis was performed to obtain the classification tree algorithm, using Gini impurity index and the Cost-Complexity pruning algorithm.[30] The algorithm allows to select the variables most likely to identify LCA-classes, which included parameters easily accessible in most outpatient settings: patient's gender, age, and BMI; assessment of nasal, bronchial and ocular symptoms, nasal impairment; and objective diagnostic tests (number of AS groups, FEV1, and FeNO). The variable importance is given by Gini index (ranging 0%-100%).[31] Additionally, we randomly divided the dataset into a training (70%) and a test set (30%) to obtain a reliable estimate of the model's predictive performance. Cohen's kappa coefficient (kappa) was used to evaluate model performance for imbalanced datasets.[32]

Details regarding LCA and CART methodology and definitions of each included respiratory/ocular symptoms are provided in Online Supplement.

3 | RESULTS

3.1 | Sample characteristics

Seven hundred and twenty-eight adults (63% female) were included, mean (SD) age of 43.9 (15.2) years, 61% were non-smokers and 11% were taking inhaled corticosteroids (ICS). Demographic and clinical characteristics of the six classes are shown in Table 1.

3.2 | Latent class analysis

A six-class model was selected as the best solution for these data, with a significantly better fitting than a five-class model ($P = 0.013$), and a non-significantly different fit from a seven-class model ($P = 0.363$) (Table S2). Furthermore, entropy of the six-class model was 0.873, a good overall certainty in classification.

Figure 1 presents the probability of latent class membership for each of the six-class LCA model and Table S3 shows the stratification of the LCA-classes, according to clinical and allergic profiles.

TABLE 1 Demographics and clinical characteristics of the 6 LCA-derived classes

	Total (n = 728)	Class 1 (n = 182; 25%)	Class 2 (n = 157; 22%)	Class 3 (n = 77 ; 11%)	Class 4 (n = 102; 14%)	Class 5 (n = 114; 16%)	Class 6 (n = 96; 13%)	P-value
Female, n (%)	461 (63)	105 (58)	89 (57)	45 (58)	65 (64)	95 (83)	62 (65)	<0.001*
Age, mean (SD)	43.9 (15.2)	46.9 (16.5)	45.5 (16.4)	39.9 (14.8)	40.3 (12.9)	45.2 (13.4)	41.3 (13.2)	<0.001**
BMI, mean (SD)	25.9 (4.7)	26.6 (4.7)	25.1 (4.0)	25.8 (4.3)	25.2 (4.6)	25.8 (5.0)	27.1 (5.3)	0.003**
Age of bronchial symptoms onset, median (P_{25}-P_{75})	8.0 (3.0-20.0)	6.0 (1.0-13.0)	3.0 (2.0-18.0)	5.0 (3.0-7.5)	7.0 (2.0-12.0)	15.0 (3.0-25.0)	12.0 (4.0-30.0)	0.06***
Smoking status, n (%)								
Non-smoker	441 (61)	108 (59)	95 (60)	47 (61)	67 (66)	75 (66)	49 (51)	0.01*
Smoker	154 (21)	47 (26)	33 (21)	20 (26)	17 (17)	22 (19)	15 (16)	
Ex-smoker	133 (18)	27 (15)	29 (18)	10 (13)	18 (18)	17 (15)	32 (33)	
Packs-year, mean (SD)	5.9 (12.8)	8.3 (15.6)	6.5 (14.2)	5.0 (10.6)	3.6 (8.4)	4.2 (8.4)	5.7 (14.0)	0.03**
Current medication, n (%)								
ICS	81 (11)	3 (2)	6 (4)	14 (18)	5 (5)	18 (16)	35 (36)	<0.001*
Number of AS groups, median (P_{25}-P_{75})	2.0 (0-6.0)	0 (0-0.7)	1.0 (0-2.0)	4.0 (3.0-5.0)	4.0 (3.0-5.0)	0 (0-0)	4.0 (3.0-6.0)	<0.001***
Level of sensitization, n (%)								
Monosensitization[a]	100 (14)	37 (21)	37 (24)	0 (0)	0 (0)	22 (19)	4 (4)	<0.001*
Polysensitization[b]	332 (46)	8 (4)	50 (32)	77 (100)	101 (100)	4 (3)	92 (96)	
Group of AS, n (%)								
Mites	336 (46)	31 (17)	55 (35)	73 (95)	77 (76)	19 (17)	81 (84)	<0.001*
Cat and dog epithelium	220 (30)	2 (1)	18 (11)	64 (83)	60 (59)	7 (6)	69 (72)	<0.001*
Molds	135 (19)	0 (0)	11 (7)	33 (43)	40 (40)	1 (1)	50 (52)	<0.001*
Pollens[c]	335 (46)	16 (9)	55 (35)	70 (91)	101 (100)	3 (3)	90 (94)	<0.001*
Lung function								
FEV_1% predicted, mean (SD)	97.4 (15.5)	99.4 (16.1)	99.6 (13.3)	95.3 (13.9)	98.9 (13.9)	97.2 (15.6)	90.1 (18.4)	<0.001**
FEV_1 <LLN, n (%)	69 (9)	16 (9)	8 (5)	6 (8)	9 (9)	8 (7)	22 (23)	<0.001*
Positive BD, n (%)	55 (8)	11 (6)	5 (3)	7 (9)	4 (4)	12 (11)	16 (17)	0.001*

AS, allergen sensitizations; BD, Bronchodilatation; BMI, body mass index; FEV1, Forced Expiratory Volume in the first second; ICS, inhaled corticosteroids; LLN, lower limit of normal; P25-P75, 25th percentile-75th percentile.

[a]Sensitization to 1 group of allergens.

[b]Sensitization to 2 or more groups of allergens.

[c]Including tree, grass and weed sensitizations.

*Chi-squared test.

**One-way analysis of variance (ANOVA).

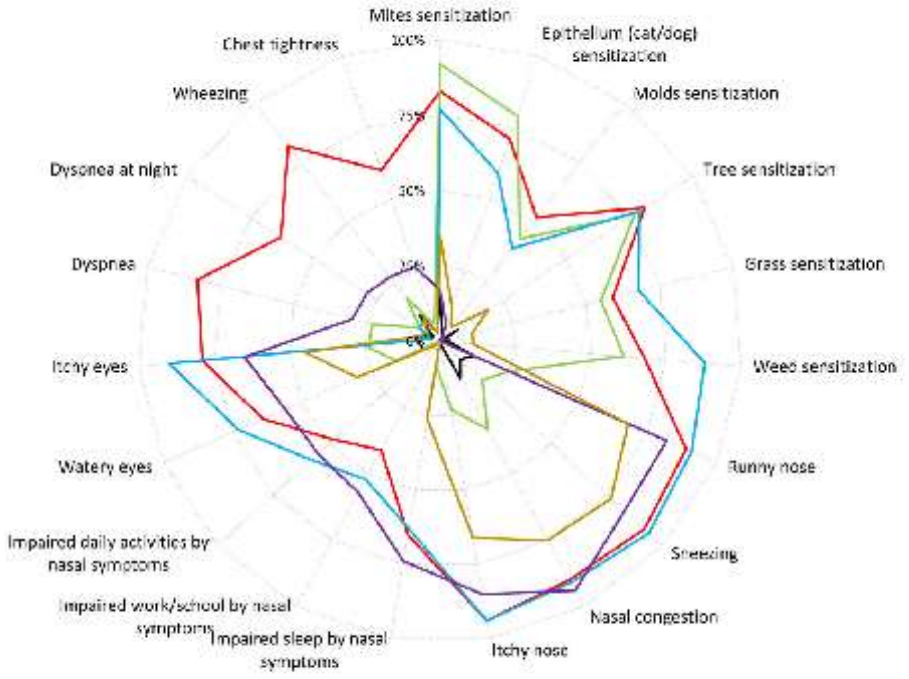
***Kruskal-Wallis test.

Average posterior probabilities were at least 89% for all classes, indicating a low chance of misclassification.

Two classes were characterized by non-allergic participants (>70% with negative SPT): classes 1 (n = 182; 25%) and 5 (n = 114; 16%). Class 1 had very low probability of having respiratory or ocular symptoms. Class 5 had a very high probability of having nasal, bronchial, and ocular symptoms with nasal severe impairment (nasal severity score≥3).

Three classes were predominantly allergic (100% sensitization): classes 3 (n = 77; 11%), 4 (n = 96; 13%), and 6 (n = 102; 14%). Class 3 had a high probability of nasal and ocular symptoms without severe nasal impairment (score≤2). Classes 4 and 6 predominantly had high nasal and ocular symptoms with nasal impairment, differing by the absence (Class 4) or presence of bronchial symptoms (Class 6).

Class 2 (n = 157; 22%) was characterized by participants with a very high probability of having nasal symptoms without severe nasal



Class 1: n = 182 (25%); Class 2: n = 157 (22%); Class 3: n = 77 (11%); Class 4: n = 102 (14%); Class 5: n = 114 (16%); Class 6: n = 96 (13%)

FIGURE 1 Proportions for the latent classes based on the estimated posterior probabilities [Colour figure can be viewed at wileyonlinelibrary.com]

impairment, with a moderately increased probability of ocular symptoms, and 55% of them were allergic.

3.3 | Latent class characteristics

There were significant differences among the six LCA-classes in all variables described in Table 1, except in age of bronchial symptoms onset ($P = 0.06$). A female predominance across all classes was observed, particularly in Class 5 (83%).

Most participants in classes 3, 4, and 6 were polysensitized, with the more frequent AS groups being: mites, pollens and cat/dog epithelia (Table 1 and Table S3). Moreover, half of the participants in Class 2 (n = 87; 55%) were sensitized, particularly to mites (n = 55; 35%) and classes 1 and 5 were mainly non-allergic.

Regarding the presence of ocular symptoms (Figure 2), the proportion of participants with ocular symptoms is significantly higher in those with severe nasal impairment (score≥3), compared to those without nasal impairment ($P < 0.001$). Similarly, the rate of participants with ocular symptoms having severe nasal impairment was significantly higher when compared to those with mild nasal impairment (score ≤2) ($P < 0.001$) or without nasal symptoms ($P < 0.001$). Also, among participants without nasal symptoms, the proportion of ocular symptoms was very low (12%).

Classes 1 and 6 represented two extreme phenotypes: Class 1 was the mildest phenotype, whereas Class 6 was the most severe, including participants with the lowest and the highest values of B-Eos, FeNO and total IgE and proportions of urgent medical care, respectively (Table 2 and Figure 3). Class 6 participants had significantly higher proportions of current use of ICS, abnormal lung function and positive bronchodilation as compared to other classes (Table 1). After Class 6, the non-allergic Class 5 had the second highest proportions of participants with current use of ICS (16%), positive bronchodilation (11%), and bronchial exacerbations, comparing to other classes.

An "Intermediate" phenotype was found. Class 3 had the lowest mean age, and when compared to other classes, had significantly higher proportions of participants with sensitization to indoor allergens, with 18% of them having current ICS medication (Table 1). Participants in Class 3 had a significantly higher proportion of bronchial exacerbations in the past 12 months, compared to classes 1 and 2 ($P < 0.001$) (Figure 3).

Serum total IgE levels and Phadiatop® sIgE were highest in sensitized groups (classes 3, 4, and 6) (Table 2). B-Eos values were

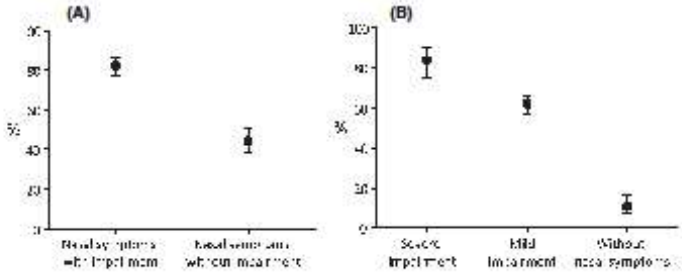


FIGURE 2 Proportions of participants with ocular symptoms, stratified by (A) the presence of nasal-related impairment and (B) ARIA severity score. The presence of nasal-related impairment defined as at least one of the following questions: impairment of daily activities by nasal symptoms, impairment of school/work by nasal symptoms and/or disturbed sleep due to nasal symptoms

significantly higher in classes 2, 4, and 6 comparing with Class 1. FeNO values were highest in classes 4 and 6, by comparison to classes 1, 3, and 5 ($P < 0.03$) (Figure S2).

Furthermore, medical diagnosis of rhinitis was common in classes 2, 4, 5, and 6 (Table S4), while asthma diagnosis was more frequent in Class 6 (75%), and less, but also prevalent, in classes 3 and 5 (35% and 31%, respectively). A high proportion of conjunctivitis was diagnosed in participants belonging to classes 4, 5, and 6. Class 1 participants had the lowest proportion of diseases medical-diagnosed, except for the proportion of other respiratory diseases (20%) (Table S4).

3.4 | CART analysis

Nine terminal nodes were formed in the classification tree (Figure 4), with a kappa(95% CI) = 0.75(0.72-0.79). Classification tree showed that ocular symptoms were the variable with the highest relative contribution to the model (37%), followed by number of AS groups (21%), having impairment of school/work by nasal symptoms (15%), presence of bronchial symptoms (13%), having impairment of daily activities (12%) and sleep disturbance due to nasal symptoms (3%).

On the right side of the tree (corresponding to participants sensitized to ≥3 AS groups), the presence of bronchial symptoms, distinguished Class 6 from all the others. When bronchial symptoms were absent, the presence/absence of impairment in daily activities and sleep disturbance by nasal symptoms, differentiated classes 3 and 4. Using a training (n = 509) and a test set (n = 219), the obtained CART algorithm was identical (Figure S3) and similar kappa was also obtained (0.73[0.67-0.80]).

4 | DISCUSSION

We identified from a general population sample using an unsupervised method six distinct classes of ARD in adults having clinically relevant differences in disease profiles. This study identified classes not determined by a clinical organ-centered diagnosis, but rather clinical characteristics and allergy profiles focusing on allergic multimorbidities. There were two non-allergic ARD phenotypes (classes 1 and 5), three allergic phenotypes (classes 3, 4, and 6), and one with half of the participants being allergic (Class 2). Moreover, we identified novel severe ARD phenotypes, with participants having concomitant ocular, nasal and bronchial symptoms (classes 5 and 6) who are prone to exacerbations. CART analysis ranked ocular symptoms as the most relevant variable to separate LCA-classes, among respiratory symptoms, nasal-related impairment and diagnostic tests readily accessible in most outpatient settings.

4.1 | Strengths and limitations of the study

Limitations of this study should be recognized. The most important is that this work lacks external validity, and therefore generalizability. A potential source of bias for unsupervised analysis is how individuals were included. In our study, this risk is very low because not only we included data in LCA that was obtained in the clinical assessment rather than data from the initial screening (self-reported), but also, more importantly, pre-screening did not comprise data regarding ocular symptoms. Although we included participants from all regions of mainland Portugal, the sensitization profile to some allergens vary in Europe and results might differ depending on geographic areas. The choice of clinical variables included in the unsupervised model may condition the obtained LCA-classes. However, their inclusion was justified not only by their clinical relevance, being currently used and validated questions comprehending the main dimensions of the ARDs,[1,13,34] but also compensating the fact that a high number of variables potentially increases the risk of some domains allocate too much weight in the modelling process.[17] The inclusion of diagnostic tests was based on parameters being easily accessible at a visit to a specialist in clinical practice and that could prove useful in disease stratification and management. A sensitivity analysis was conducted using multiple imputation of incomplete data[35] and the obtained LCA-classes were similar when imputing all missing values (data not shown). Finally, the cross-sectional design of this study does not allow us to address latent classes stability over time.

TABLE 2 Distribution and comparison of FeNO and laboratory tests, according to each obtained LCA-classes

	Class 1 (n = 182; 25%)	Class 2 (n = 157; 22%)	Class 3 (n = 77; 11%)	Class 4 (n = 102; 14%)	Class 5 (n = 114; 16%)	Class 6 (n = 96; 13%)	Significant P-values*
FeNO, geom mean (CI 95%)	17.9 (16.0-19.9)	18.1 (16.2-20.0)	24.9 (18.9-30.9)	26.5 (22.0-31.0)	19.3 (16.1-22.4)	35.1 (28.7-41.5)	C1/C4, C1/C6, C3/C4, C3/C6, C4/C5, C5/C6
S-ECP, geom mean (CI 95%)	13.2 (11.2-15.2)	14.5 (11.9-17.1)	13.2 (10.6-15.9)	15.1 (11.7-18.5)	12.0 (9.8-14.3)	16.8 (13.7-20.0)	N.S.
B-Eos, geom mean (CI 95%)	158.7 (133.6-183.8)	177.2 (152.3-202.0)	200.9 (169.0-232.7)	205.3 (174.7-236.0)	189.3 (161.8-216.8)	237.2 (197.4-277.1)	C1/C2, C1/C4, C1/C6
Phadiatop® sIgE, median (P_{25}-P_{75})	0 (0-0)	0 (0-1.6)	3.6 (0-16.5)	7.4 (1.9-23.2)	0 (0-0)	13.2 (0.6-31.3)	C1/C3, C1/C4, C1/C6, C2/C3, C2/C4, C2/C6, C3/C5, C4/C5, C5/C6
Total IgE, geom mean (CI 95%)	85.5 (52.4-118.7)	76.9 (54.7-99.2)	202.8 (138.5-267.0)	184.8 (111.0-258.6)	52.2 (36.2-68.2)	227.6 (170.0-285.3)	C1/C3, C1/C4, C1/C6, C2/C3, C2/C4, C2/C6, C3/C5, C4/C5, C5/C6

B-Eos, blood eosinophilic count; C, class; CI, confidence interval; FeNO, fractional exhaled nitric oxide; N.S., non-significant; S-ECP, serum eosinophilic cationic protein; sIgE, specific IgE.
*Significant if P-values<0.05, using Kruskal-Wallis test with Bonferroni correction.

The unsupervised analysis based on airways symptoms and allergy profile data remains a powerful approach toward ARD phenotyping.[10,16,36] In our study, we not only identified data-driven phenotypes but also, we provided information on the relevance of the variables that better distinguish the LCA-classes, using the supervised method CART analysis. While some previous studies used only unsupervised-clustering methods to identify phenotypes of ARD,[13-14] the present study extended this approach into providing information on the importance of the variables that best distinguish between the obtained classes, using CART analysis. To our knowledge, this is the first time CART analysis is studied in a population-based study of ARD. This analysis has various advantages over other methods, including multivariable logistic regression: it is a nonparametric method; results are summarized in a tree, much simpler to interpret and more practical in a clinical setting; and measures the variable relevance in the model (relative impact of the predictors on the output).[35] Also, our tree algorithm based on parameters already used in clinical practice performed well using a training and a test set, suggesting a high potential to be further applied.

A major strength of the study is that we used ocular symptoms independently for the first time in LCA and they were found to be extremely informative.

4.2 | Interpretation of the results

4.2.1 | Profiles differed by their association with IgE allergy

There were two non-allergic ARD phenotypes (classes 1 and 5), three allergic phenotypes (classes 3, 4, and 6) and one with 50% participants being allergic (Class 2). Class 1 was the mildest phenotype, without symptoms or inflammation. Class 5 comprised predominantly females, with low FeNO and B-Eos, and a high proportion of urgent health care use, suggesting that bronchial symptoms can be linked with non-allergic rhinitis[37] or rhinosinusitis. Moreover, Class 5 participants have some characteristics similar to the non-allergic rhinitis phenotype obtained by cluster analysis in other studies.[13,14]

4.2.2 | Multimorbidity is a necessary component of ARD

In our study, the unsupervised analysis did not identify clusters of participants having asthma/rhinitis/conjunctivitis only, supporting the concept of "one airway one disease"[38] and considering the view that rhinitis, conjunctivitis, and asthma are different manifestations of the same disease. Particularly, there was no asthma cluster without rhinitis. This accords with epidemiologic studies such as ECRHS where asthma alone represents <10% of the asthmatic population. Thus, it seems that in clinical practice, multimorbidity should always be investigated in ARD patients. This fact also holds for a precision medicine approach based on treatable traits, rather than diagnostic labels, in the clinical management of the ARDs.[1,8,39] One surprising issue is the prevalence of conjunctivitis in four classes. However, the diagnosis was made by a physician.

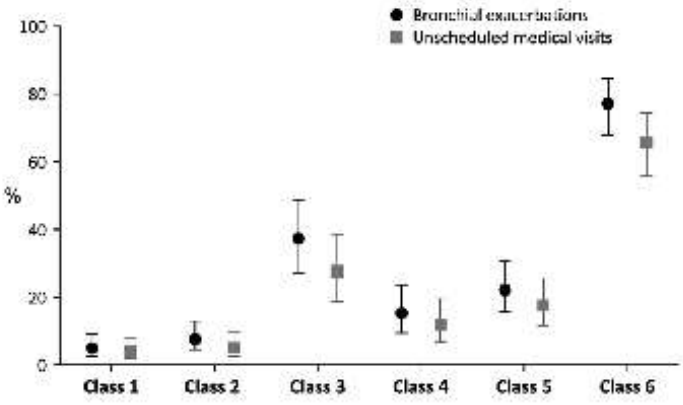


FIGURE 3 Proportions of participants with bronchial exacerbations and unscheduled medical visits, stratified by LCA-classes

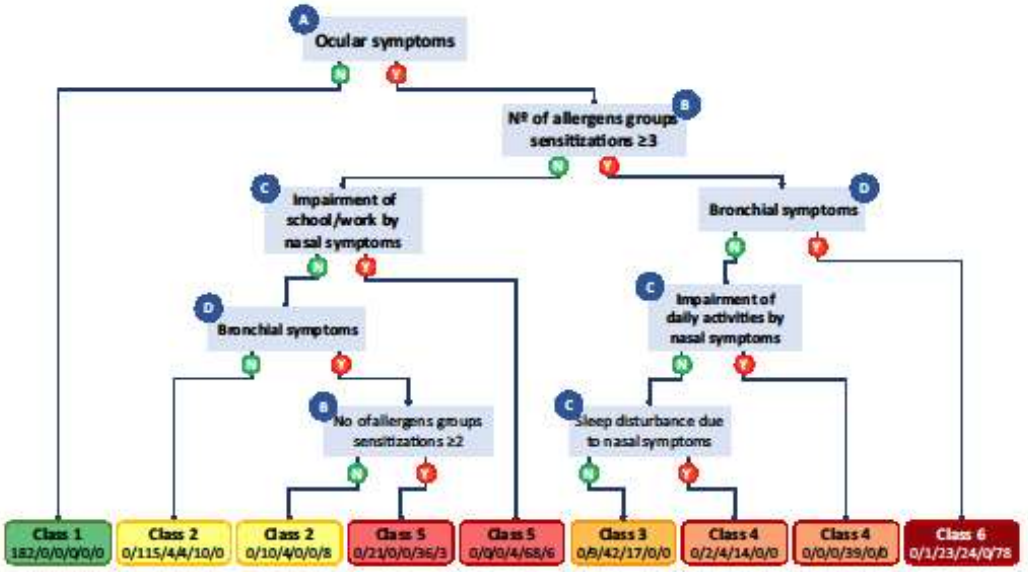


FIGURE 4 Classification tree algorithm generated by CART analysis using easily accessible parameters available in the clinical setting. (A) At least 1 ocular symptom: Watery eyes and/or Itchy eyes. (B) Mites, Cat/dog epitheliums, Tree, Grass, Weed, Molds. (C) Impairment due to nasal symptoms: Sneezing, Rhinorrhea, Nasal pruritus, Nasal congestion; (D) At least one bronchial symptom: Dyspnea, Dyspnea at night, Wheezing and/or Chest tightness. N: No; Y: Yes [Colour figure can be viewed at wileyonlinelibrary.com]

The obtained LCA-classes found in our sample, derived from the general population, were very similar to previously published clinical phenotypes in non-allergic and allergic rhinoconjunctivitis patients.[40,41] However, we not only reinforced the different patterns of multimorbidity in participants with rhinitis, but also, we identified a specific class of participants polysensitized predominantly to indoor allergens (Class 3), with mild nose and ocular symptoms but presenting bronchial exacerbations, suggesting that different phenotypes of ARD can be obtained when analyzing a more comprehensive sample (population-based) and not derived from a secondary healthcare setting.

4.2.3 | Ocular symptoms are essential to identify clusters of ARD patients

The inclusion of ocular symptoms, independently from other symptoms was not explored in cluster analyses except by MeDALL,[56] but, in this study, only rhinoconjunctivitis was considered. We observed that the proportion of ocular symptoms was significantly higher in participants with severe ARD (Classes 3 and 4 [severe nasal disease], 5 and 6 [severe nasal and bronchial disease prone to exacerbations]) and were less expressed in participants with milder ARD (classes 2 and 3). Finally, it

was found that ocular symptoms are associated with the severity of nasal symptoms suggesting that rhinitis and rhinoconjunctivitis represent two distinct phenotypes. This accords with a paper showing that the number of allergens recognized in the two phenotypes differ (low number in rhinitis, significantly higher number for rhinoconjunctivitis, Sirouxet al, submitted).

The polysensitized multimorbid phenotype (Class 6) proposed by MeDALL[4,5] was confirmed and found to be associated with rhinitis and asthma severity. Moreover, this study also identified conjunctivitis as another component of the multimorbid ARD phenotype. This class has some characteristics similar to those found in other unsupervised clustering studies, which has labeled as "late-onset, inflammation predominant".[11,12,42]

4.2.4 | CART analysis ranked ocular symptoms as the most relevant variable to separate LCA-classes

The tree algorithm not only reinforced the importance of including the presence of ocular symptoms in the expression of ARD phenotypes, among other easily accessible parameters used in a routine clinical practice, but also it may help to classify patients in clinical settings and to distinguish between lower or higher degrees of airways allergic multimorbidity.

4.2.5 | This study has an important impact in epidemiology

The concept of the importance of ocular symptoms in allergic multimorbidity was raised by a big data study using an App in uncharacterized users.[43] This hypothesis-generating study needed to be confirmed by a classical epidemiologic study and the present study is the first to show the relevance of big data in the future of epidemiology for respiratory and allergic diseases.

4.2.6 | The characterization of airway inflammation in ARD is also essential[44]

In our data, we observed that B-Eos values were lower in participants of class 1, separated those with and without nasal and ocular symptoms, and lower FeNO values in non-allergic participants and without nasal-related impairment, differentiating phenotypes with different allergy profiles and symptoms severity, supporting the view that the combination of these two biomarkers should be used for personalized treatment.[45,46]

4.3 | Generalizability

This study performed in Portugal needs to be confirmed in other countries as regional variations exist in allergens. However, a study on multimorbidity and allergen sensitizations showed that similar data were observed in France and Sweden suggesting common biological mechanisms.[47] After validation of these phenotypes in other populations, further work should address biologic associations, such as total IgE, sIgE, B-Eos, S-ECP, and mechanisms within these phenotypes. Also, studies on the differentiation of allergic vs non-allergic conjunctivitis, and regarding symptoms' deterioration among non-smokers vs smoker/former-smokers and among different sensitizations, should be done, especially with prospective designs.

5 | CONCLUSION

The findings of our study challenge the conventional disease classification of a "classical" clinical diagnosis organ-based approach, to a combination of unsupervised and supervised analysis bringing novel insights to phenotyping ARD while helping confirm and integrate phenotypes previously reported. Six latent classes of ARD were identified in adults from a general population sample, with clinically relevant differences in disease profiles. Novel severe phenotypes of participants with co-occurrence of ocular, nasal and bronchial symptoms and exacerbation-prone, were revealed. Furthermore, the tree algorithm not only reinforced the importance of ocular symptoms in the expression of ARD phenotypes but also may help to distinguish between lower and higher degree of allergic airway multimorbidity.

ACKNOWLEDGMENTS

First author is supported by a PhD Grant (PD/BD/113659/2015), financed by the European Social Fund and national funds of MCTES (Ministério da Ciência, Tecnologia e Ensino Superior) through FCT (Fundação para a Ciência e Tecnologia, I.P) PhD Programme ref PD/0003/2013 - Doctoral Programme in Clinical and Health Services Research (PDICSS). The data collection was part of a funded FCT project (PTDC/SAU-SAP/ 119192/2010). The funding source had no role in the study design, the data collection, analysis and interpretation, or the writing of the manuscript.

CONFLICTS OF INTEREST

Dr. Amaral has nothing to disclose; Dr. Bousquet reports personal fees and other from Chiesi, Cipla, Hikma, Menarini, Mundipharma, Mylan, Novartis, Sanofi-Aventis, Takeda, Teva, Uriach, other from KYomed-INNOV, outside the submitted work; Dr. Pereira has nothing to disclose; Dr. Araújo has nothing to disclose; Dr. Sá-Sousa has nothing to disclose; Dr. Jacinto has nothing to disclose; Dr. Almeida has nothing to disclose; Dr. Delgado has nothing to disclose; Dr. Fonseca has nothing to disclose.

AUTHOR CONTRIBUTIONS

RA, JB, LD, and JAF contributed to the conception and design of the work, analysis and interpretation of data, drafting and revising the manuscript. RA, AMP, LA, ASS, RAI, and TJ contributed to data collection, analysis, and interpretation, as well as revising the manuscript. All authors read and approved the final version of the manuscript.

ORCID

Rita Amaral https://orcid.org/0000-0002-0233-830X

REFERENCES

1. Bousquet J, van CP, Khaltaev N. Allergic rhinitis and its impact on asthma. *J Allergy Clin Immunol*. 2001;108:S147-S334.
2. Pinart M, Benet M, Annesi-Maesano I, et al. Comorbidity of eczema, rhinitis, and asthma in IgE-sensitised and non-IgE-sensitised children in MeDALL: a population-based cohort study. *Lancet Respir Med*. 2014;2:131-140.
3. Johansson SG, Bieber T, Dahl R, et al. Revised nomenclature for allergy for global use: report of the Nomenclature Review Committee of the World Allergy Organization, October 2003. *J Allergy Clin Immunol*. 2004;113:832-836.
4. Anto JM, Bousquet J, Akdis M, et al. Mechanisms of the Development of Allergy (MeDALL): introducing novel concepts in allergy phenotypes. *J Allergy Clin Immunol*. 2017;139:388-399.
5. Bousquet J, Anto JM, Wickman M, et al. Are allergic multimorbidities and IgE polysensitization associated with the persistence or re-occurrence of foetal type 2 signalling? The MeDALL hypothesis *Allergy*. 2015;70:1062-1078.
6. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet*. 2012;380(9836):37-43.
7. Hodson R. Precision medicine. *Nature*. 2016;537(7619):S49.
8. Agusti A, Bel E, Thomas M, et al. Treatable traits: toward precision medicine of chronic airway diseases. *Eur Respir J*. 2016;47:410-419.
9. Bousquet J, Anto JM, Sterk PJ, et al. Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med*. 2011;3:43.
10. Bousquet J, Anto JM, Akdis M, et al. Paving the way of systems biology and precision medicine in allergic diseases: the MeDALL success story. *Allergy*. 2016;71:1513-1525.
11. Siroux V, Basagana X, Boudier A, et al. Identifying adult asthma phenotypes using a clustering approach. *Eur Respir J*. 2011;38:310-317.
12. Haldar P, Pavord ID, Shaw DE, et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med*. 2008;178:218-224.
13. Burte E, Bousquet J, Varraso R, et al. Characterization of rhinitis according to the asthma status in adults using an unsupervised approach in the EGEA study. *PLoS One*. 2015;10:e0136191.
14. Kurukulaaratchy RJ, Zhang H, Patil V, et al. Identifying the heterogeneity of young adult rhinitis through cluster analysis in the Isle of Wight birth cohort. *J Allergy Clin Immunol*. 2015;135:143-150.
15. Lee E, Lee SH, Kwon JW, et al. Atopic dermatitis phenotype with early onset and high serum IL-13 is linked to the new development of bronchial hyperresponsiveness in school children. *Allergy*. 2016;71:692-700.
16. García-Aymerich J, Benet M, Saeys Y, et al. Phenotyping asthma, rhinitis and eczema in MeDALL population-based birth cohorts: an allergic comorbidity cluster. *Allergy*. 2015;70:973-984.
17. Collins LM, Lanza ST. The latent class model. In: Collins LM, Lanza ST, eds. *Latent Class and Latent Transition Analysis*. Hoboken, NJ: Wiley; 2010:23-47.
18. Moore WC, Meyers DA, Wenzel SE, et al. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *Am J Respir Crit Care Med*. 2010;181:315-323.
19. Marshall RJ. The use of classification and regression trees in clinical epidemiology. *J Clin Epidemiol*. 2001;54:603-609.
20. Bousquet J, Arnavielhe S, Bedbrook A, et al. The Allergic Rhinitis and its Impact on Asthma (ARIA) score of allergic rhinitis using mobile technology correlates with quality of life: the MASK study. *Allergy*. 2018;73:505-510.
21. Sa-Sousa A, Pereira A, Almeida R, et al. Adult asthma scores - development and validation of multivariable scores to identify asthma in surveys. *J Allergy Clin Immunol Pract*. 2018; pii: S2213-2198(18)30428-8. https://doi.org/10.1016/j.jaip.2018.06.024. [Epub ahead of print].
22. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61:344-349.
23. Bousquet J, Heinzerling L, Bachert C, et al. Practical guide to skin prick tests in allergy to aeroallergens. *Allergy*. 2012;67:18-24.
24. Heinzerling L, Frew AJ, Bindslev-Jensen C, et al. Standard skin prick testing and sensitization to inhalant allergens across Europe - a survey from the GA2LEN network. *Allergy*. 2005;60:1287-1300.
25. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med*. 1999;159:179-187.
26. Miller MR. Standardisation of spirometry. *Eur Respir J*. 2005;26:319-338.
27. Quanjer PH, Cole TJ, Hall GL, Culver BH. Multi-ethnic reference values for spirometry for the 3-95 years age range: the global lung function 2012 equations. *Eur Respir J*. 2013;40:1324-1343.
28. American Thoracic S, European Respiratory S. ATS/ERS recommendations for standardized procedures for the online and offline measurement of exhaled lower respiratory nitric oxide and nasal nitric oxide, 2005. *Am J Respir Crit Care Med*. 2005;171:912-930.
29. Wang J, Wang X. Mixture modeling. In: Balding DJ, ed. *Structural Equation Modeling: Applications using Mplus*. New York: Wiley; 2012: pp 289-390.
30. Witten IH, Frank E, Hall MS. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, VT: Elsevier; 2005.
31. Sauve M, Tuleau-Malot C. Variable selection through CART. *ESAIM Probab Stat*. 2014;18:770-798.
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
33. Fonseca JA, Nogueira-Silva L, Morais-Almeida M, et al. Control of Allergic Rhinitis and Asthma Test (CARAT) can be used to assess individual patients over time. *Clin Transl Allergy*. 2012;2:16.
34. Nguyen JM, Holbrook JT, Wei CY, Gerald LB, Teague WG, Wise RA. Validation and psychometric properties of the Asthma Control Questionnaire among children. *J Allergy Clin Immunol*. 2014;133:91-97.
35. Rubin DB. Multiple imputation. In: Rubin DB, ed. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons Inc; 1987:1-23.
36. Bousquet J, Anto J, Auffray C, et al. MeDALL (Mechanisms of the Development of ALLergy): an integrated approach from phenotypes to systems medicine. *Allergy*. 2011;66:596-604.
37. Shaaban R, Zureik M, Soussan D, et al. Rhinitis and onset of asthma: a longitudinal population-based study. *Lancet*. 2008;372(9643):1049-1057.
38. Grossman J. One airway, one disease. *Chest*. 1997;111(2 Suppl):11S-16S.
39. Bousquet J, Vignola AM, Demoly P. Links between rhinitis and asthma. *Allergy*. 2003;58:691-706.
40. Di Lorenzo G, Pacor ML, Amodio E, et al. Differences and similarities between allergic and nonallergic rhinitis in a large sample of adult patients with rhinitis symptoms. *Int Arch Allergy Immunol*. 2011;155:263-270.
41. Molgaard E, Thomsen SF, Lund T, Pedersen L, Nolte H, Backer V. Differences between allergic and nonallergic rhinitis in a large sample of adolescents and adults. *Allergy*. 2007;62:1033-1037.
42. de Vries R, Dagelet YWF, Spoor P, et al. Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. *Eur Respir J*. 2018;51:1701817.
43. Bousquet J, Devillier P, Anto JM, et al. Daily allergic multimorbidity in rhinitis using mobile technology: a novel concept of the MASK study. *Allergy*. 2018;73:1622-1631.

44. Jeffery PK, Haahtela T. Allergic rhinitis and asthma: inflammation in a one-airway condition. *BMC Pulm Med*. 2006;6(Suppl 1):S5.
45. Malinovschi A, Fonseca JA, Jacinto T, Alving K, Janson C. Exhaled nitric oxide levels and blood eosinophil counts independently associate with wheeze and asthma events in National Health and Nutrition Examination Survey subjects. *J Allergy Clin Immunol*. 2013;132:821-827.
46. Malinovschi A, Janson C, Borres M, Alving K. Simultaneously increased fraction of exhaled nitric oxide levels and blood eosinophil counts relate to increased asthma morbidity. *J Allergy Clin Immunol*. 2016;138:1301-1308.
47. Siroux V, Ballardini N, Soler M, et al. The asthma-rhinitis multimorbidity is associated with IgE polysensitization in adolescents and adults. *Allergy*. 2018;73:1447-1458.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Amaral R, Bousquet J, Pereira AM, et al. Disentangling the heterogeneity of allergic respiratory diseases by latent class analysis reveals novel phenotypes. *Allergy*. 2019;74:698–708. https://doi.org/10.1111/all.13670