Ascertaining Patterns of Asthma Symptoms in Childhood using Machine Learning Methods

A thesis submitted to The University of Manchester for the degree of

Doctor of Philosophy

In the Faculty of Biology, Medicine and Health.

2019

Matea Deliu

School of Health Sciences
Division of Informatics, Imaging and Data Science

Blank page

Table of Contents

Final word count (including footnotes, endnotes, references): 53640

List of Tables

List of Figures

Abstract

The prevalence and incidence of asthma in children is continually rising and creating a burden on health systems due to high health care costs. Asthma is a heterogeneous disease however current definitions do not capture the heterogeneity of this complex condition as it is becoming increasingly clear that it is not a single disease but rather a collection of syndromes which consist of a number of disease subtypes ('endotypes') with similar observable and measurable clinical characteristics ('phenotypes'). Identifying true endotypes of asthma and disaggregating the heterogeneity of the disease is required for achieving better pathophysiological mechanism-based treatment targeting, and thus delivering genuinely personalised pharmacological treatment in asthma.

Methods of ascertaining these endotypes have ranged from investigator-led pattern identification in the clinical setting, to supervised and unsupervised statistical modelling techniques that utilize large scale data and computer algorithms to find the latent (hidden, unknown *a-priori*) patterns of observable features (such as symptoms, medication use, allergic sensitization, lung function). Data-driven approaches allow the data to essentially speak for itself without any *a-priori* hypotheses imposition guiding the analysis. This ultimately eliminates investigator bias and enables novel hypotheses to be generated.

Using two different rich data sources (Turkish population cohort and Manchester Asthma and Allergy Study) both cross-sectionally and longitudinally, this thesis aimed to answer the following research questions: 1) Can we use data-driven methods to uncover patterns among asthma datasets and how can this help guide our further understanding of the disease? 2) What main features of the asthma syndrome can be used to ascertain the heterogeneity of the disease? 3) How can we exploit the wealth of data provided by longitudinal birth cohorts in order to understand the severity of asthma?

Chapter 2 of the thesis introduced and explained in detail the use of machine learning methods such as cluster analysis and latent class analysis that have been increasingly frequently used in ascertaining patterns of asthma phenotypes. Chapter 3 then puts this data-driven methodology in context by discussing the advancements in knowledge acquired from the use of these algorithms. Using cross-sectional data from Turkey, Chapter

4 creates a framework for the discovery of stable and clinically meaningful asthma subtypes by blending data with clinical expert domain knowledge to identify four main informative features (age of onset, atopy, exacerbations, severity). To that end, Chapters 5 and 6 used longitudinal data in order to explore exacerbations and asthma severity in more detail. Two independent exacerbation subtypes were identified (frequent and infrequent exacerbations) along with three wheeze severity states (mild/moderate wheeze, severe wheeze, and transitioning wheeze).

This thesis represents an advancement on our current knowledge of the heterogeneity of asthma by identifying novel results through the use of machine learning methodologies.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

This PhD thesis would not have been possible without the amazing help and support of colleagues, friends and family, whom deserve to be acknowledged.

Firstly, to my supervisors: Prof. Adnan Custovic, Dr. Nophar Geifman, Dr. Matthew Sperrin. Your devotion, guidance, and knowledge has been paramount in both the completion of the PhD and in me developing skills to become a clinician/health data scientist/informatician. Without your patience, none of this would have been possible. Thank you for the time you have taken to supervise me over the past four years and I look forward to ongoing future collaborations with you all.

Secondly, to my parents and fiancée: you have been with me through the tough times, constantly pushing me to keep going. Thank you for your continued love and support which has undeniably contributed to me completing this PhD.

Finally, I'd like to thank all the students/researchers/staff within both the Health e-Research Centre, University of Manchester, and at Imperial College London (with particular mention to Sara Fontanella and Sadia Haider) who I have had utmost pleasure in working alongside with.

Thank you to everyone,

Matea

About the Author

1.1 Candidate Degrees

2008-2014 MD (Medical Doctor), University of Zagreb School of Medicine

1.2 Research Interests

Matea Deliu is a qualified medical doctor interested in utilising different machine learning methods in order to find patterns among asthma datasets leading to better understanding of asthma endotypes.

1.3 Publications

Published peer-reviewed papers arising directly from this PhD include:

Features of asthma which provide meaningful insights into understanding the disease heterogeneity
Matea Deliu, S. Tolga Yavuz, Matthew Sperrin, Danielle Belgrave, Umit M. Sahiner, Cansin Sackesen, Adnan Custovic, Omer Kalayci
Clin Exp Allergy. 2017 Aug 22. doi: 10.1111/cea.13014

Asthma Phenotypes in Childhood
Matea Deliu, Danielle Belgrave, Matthew Sperrin, Iain Buchan, Adnan Custovic
Expert Rev Clin Immunol. 2017 Jul;13(7):705-713. doi: 10.1080/1744666X.2017.1257940.

Identification of Asthma Subtypes Using Clustering Methodologies
Matea Deliu, Matthew Sperrin, Danielle Belgrave, Adnan Custovic
Pulmonary Therapy 2016, 2(1), 19-41 DOI: 10.1007/s41030-016-0017-z

Blank page

Chapter 1 General introduction

1.1 Background and Rationale for Thesis

1.1.2 What is Machine Learning?

Machine learning is a data-driven class of algorithms with many applications in data mining. Machine learning methods work by uncovering hidden relationships between datasets by using models that classify or predict particular outcomes.[1] These algorithms learn exclusively from data and so are able to adapt to new data being introduced. The most common methods use supervised learning and unsupervised learning.

Supervised learning

In supervised learning, the labels (or inputs and outputs of an algorithm) are known and the aim is to predict/model them. For example, in face recognition, a supervised learning algorithm would learn what a face is and what a face is not based on multiple images tagged "face-yes and face-no". Once that is learned, it would be able to predict whether a new image is a face or not. Examples of supervised methods are: decision trees, random forests, support vector machines.

Unsupervised learning

In unsupervised learning, no labels are given and the model explores the data to differentiate underlying structures and connections. Referring to the same face recognition example, the model would not know what a face is, but would be able to group (cluster) different images based on characteristics (i.e. faces vs animals vs balls, etc). Clustering methods are the most common and will be used exclusively in this PhD thesis.

1.1.3 Machine Learning and Utilisation in Understanding Asthma

Big data and its utilisation through machine learning has been finding niches in all aspects of modern life. Companies like Google, Facebook, Netflix are using algorithms to learn aspects of our day to day life in order to make improvements in their service. Over the last decade, the application of machine learning to health care data has been revolutionising the way we think and understand different disease mechanisms. Recent examples have demonstrated that data driven computer algorithms can perform on par with clinical decision making.[2,3]

Machine learning methods have been gaining recognition in helping us understand the heterogeneity of asthma – a chronic disease inducing narrowing of small airways that creates great burden on both the patient and the health care system. Within this realm, we are inundated with a vast array of datasets, investigator and data led analyses that have provided answers to different questions. However, there have not been any unifying methodologies and so many results have been inconclusive. Machine learning can provide an unbiased data-driven method to ascertaining patterns among datasets as asthma is an extremely heterogeneous condition – an umbrella term for several diseases manifesting with common symptoms such as wheeze, cough, breathlessness, or chest tightness, but differing in aetiology, pathophysiologic mechanisms, and treatment response.[4-6] The term 'asthma phenotype' has been has been frequently used to describe observable characteristics and disentangling these different types will help us understand the underlying functional or pathophysiological mechanisms, genetics, and environmental factors, otherwise known as 'asthma endotypes' or subtypes.[7] The aim of using data-driven approaches is to avoid pre-established hypotheses and better classify patients into these subtypes as they will share some common features. By using the wealth and quality of data provided to us by longitudinal birth cohorts, machine learning approaches can learn from the many patterns of the natural history of asthma in order to apply generalisations of the developmental profiles of the syndrome to a wider population. By utilising these statistical techniques, and taking into account between cohort heterogeneity, we can extrapolate endotypes with similar characteristics across wider populations in order to identify a more complete picture of the asthma syndrome.

The rationale for this is that we are seeing varying levels of treatment response from patients and so by identifying these subtypes, we would be able to move away from a unified “one size fits all” approach to a targeted personalised or stratified therapeutic approach.[8] The clinical relevance of this is that we are yet to understand which currently available treatment strategies utilised in children are beneficial in the long term.[9,10] Therefore by targeting childhood, the aim would be to prevent the development or further progression into adulthood. However, current research has merely scratched the surface of this underlying complex problem and much remains to be done in moving from this current hypothetical construct to application in everyday clinical practice.

1.2 Research Questions and Thesis Structure

This thesis is structured in “journal format” (previously called “alternative format”), as a series of papers accepted in, pending submission, or already submitted to, peer-reviewed journals. It is aimed at answering the following research questions:

1) Can we use data-driven methods to uncover patterns among asthma datasets and how can this help guide our further understanding of the disease?
2) What main features of the asthma syndrome can be used to ascertain the heterogeneity of the disease?
3) How can we exploit the wealth of data provided by longitudinal birth cohorts in order to understand the severity of asthma?

The papers are arranged in a way that tries to answer each research question in sequential order. The thesis then concludes with a general discussion drawing the research together and thereby providing ideas for future work.

1.3 Author Contributions

As described by The University of Manchester guidance on journal format thesis presentations, the contributions made by each author to the herein published papers are given below.

- Chapter 2: Identification of Asthma Subtypes using Clustering Methodologies. *Pulmonary Therapy 2016, 2(1), 19-41 DOI: 10.1007/s41030-016-0017-z.*

  - MD facilitated the literature review and the organisation of the paper as well as writing the initial draft of the manuscript. MS, DB, AC critically reviewed the manuscript.

- Chapter 3: Asthma Phenotypes in Childhood. *Expert Rev Clin Immunol. 2017 Jul;13(7):705-713. doi: 10.1080/1744666X.2017.1257940.*
  - MD facilitated the literature review and the organisation of the paper as well as writing the initial draft of the manuscript. MS, DB, AC critically reviewed the manuscript.
- Chapter 4: Features of Asthma Which Provide Meaningful Insights Into Understanding the Disease Heterogeneity. *Clin Exp Allergy. 2017 Aug 22. doi: 10.1111/cea.13014*
  - MD, OK, and AC designed the study. TY collected all the data. MD conducted the analysis, interpreted the findings with AC and MS, and wrote the initial manuscript draft. All other writers reviewed the manuscript for intellectual content.
- Chapter 5: Longitudinal Patterns of Asthma Exacerbations from Infancy to School Age. *Clin Exp Allergy (Under review)*
  - MD and AC designed the study. MD conducted the analysis, interpreted the findings with AC and MS, NG, SF, SH, and wrote the initial manuscript draft. All writers reviewed the manuscript for intellectual content.
- Chapter 6: Patterns of wheeze severity from early childhood to late adolescence: Longitudinal transition analysis in a birth cohort study. *Pending submission*
  - MD and AC designed the study. SF and MD conducted the analysis. MD interpreted the analysis with AC, and SF and wrote the initial manuscript draft. All writers reviewed the manuscript for intellectual content.

1.4 References

1. Yoo C, Ramirez L, Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine. *Int Neurourol J* 2014; **18**(2): 50-7.

2. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 2016; **316**(22): 2402-10.

3. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. *Nature* 2018; **555**(7697): 487-92.

4. Wenzel SE. Asthma: defining of the persistent adult phenotypes. *Lancet* 2006; **368**(9537): 804-13.

5. Smith JA, Drake R, Simpson A, Woodcock A, Pickles A, Custovic A. Dimensions of respiratory symptoms in preschool children: population-based birth cohort study. *Am J Respir Crit Care Med* 2008; **177**(12): 1358-63.

6. Papadopoulos NG, Arakawa H, Carlsen KH, et al. International consensus on (ICON) pediatric asthma. *Allergy* 2012; **67**(8): 976-97.

7. Lotvall J, Akdis CA, Bacharier LB, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol* 2011; **127**(2): 355-60.

8. Stein RT, Martinez FD. Asthma phenotypes in childhood: lessons from an epidemiological approach. *Paediatr Respir Rev* 2004; **5**(2): 155-61.

9. Brand PL, Baraldi E, Bisgaard H, et al. Definition, assessment and treatment of wheezing disorders in preschool children: an evidence-based approach. *Eur Respir J* 2008; **32**(4): 1096-110.

10. van Aalderen WM, Sprikkelman AB. Inhaled corticosteroids in childhood asthma: the story continues. *Eur J Pediatr* 2011; **170**(6): 709-18.

Chapter 2 Identification of asthma subtypes using clustering methodologies

Matea Deliu; Matthew Sperrin; Danielle Belgrave; Adnan Custovic

Pulmonary Therapy 2016, 2(1), 19-41 DOI: 10.1007/s41030-016-0017-z

1.1 Abstract

Asthma is a heterogeneous disease comprising a number of subtypes which may be caused by different pathophysiologic mechanisms (sometimes referred to as endotypes), but may share similar observed characteristics (phenotypes). The use of unsupervised clustering in adult and paediatric populations has identified subtypes of asthma based on observable characteristics such as symptoms, lung function, atopy, eosinophilia, obesity, and age of onset. Here we describe different clustering methods and demonstrate their contributions to our understanding of the spectrum of asthma syndrome. Precise identification of asthma subtypes and their pathophysiological mechanisms may lead to stratification of patients with more precise therapeutic and prevention approaches.

2.1 Introduction

Asthma is a heterogeneous disease, and the most recent Global Strategy for Asthma Management and Prevention (GINA) consensus defines it as a condition characterised by the presence of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable airflow obstruction [1]. However, various definitions of asthma do not capture the heterogeneity of this common complex condition. It is becoming increasingly clear that asthma is not a single disease, but a syndrome which consists of a number of disease subtypes with similar observable clinical characteristics[2]. These observable characteristics of the disease are often referred to as asthma phenotypes. The term asthma endotype is not synonymous with phenotype, and it should be used to refer to the distinct disease entity under the umbrella diagnosis of asthma, which has defined pathophysiological mechanisms that give rise to clinical symptoms[3]. It is worth emphasising that the same observable characteristic (i.e. phenotype) can arise as a consequence of different underlying pathologies (i.e. endotypes), which is consistent with observations showing that there are subtypes of asthma that share similar clinical symptoms, but have differing underlying pathophysiological mechanisms[4]. There are numerous examples in other disease areas of a similar or identical clinical presentation arising as a consequence of different pathology (for example, fever in childhood can be caused by numerous different mechanisms).

The traditional constructs of "asthma phenotypes" have been largely descriptive with little uniformity, and were usually informed by subjective observations of single dimensions of the disease, such as triggering factors (e.g. extrinsic and intrinsic asthma[5], exercise-induced asthma[6]), patterns of airway obstruction (e.g. reversible and irreversible asthma[7]), or pathology (e.g. eosinophilic and non-eosinophilic asthma[8]).

In paediatric asthma, change over time in symptoms such as wheeze has been used to define phenotypes of wheezing illness during childhood[9]. For example, based on the clinical observation about changes in the temporal pattern of wheezing illness during childhood which was confirmed in the birth cohort study (Tucson Children's Respiratory Study), Martinez *et al* divided children into three groups (or phenotypes) of wheezing - transient early wheezers, late-onset wheezers and persistent wheezers[10]. Although these phenotypes are clinically meaningful in their association with lung function and subsequent development of asthma[11], their distinct underlying pathophysiological mechanisms have not been elucidated and confirmed – i.e., they cannot be considered as endotypes.

Based on the expert opinion and consensus, Lotvall *et al* [4] suggested the existence of six asthma endotypes: aspirin-sensitive asthma, allergic broncho-pulmonary mycosis, allergic asthma, asthma predictive index-positive preschool wheezers, severe late-onset hypereosinophilic asthma and asthma in cross-country skiers. However, the well-defined pathophysiological mechanisms and biomarkers which differentiate between these proposed endotypes have not been discovered, and there is no universal agreement that these subtypes of asthma represent true endotypes[12]. At this time, the endotype concept remains largely hypothetical, but may have a tangible value in helping us to formulate strategies to better understand the mechanisms underlying different asthma-related diseases and, through this, identify more effective stratified treatment strategies[13].

In recent years, approaches to subtyping asthma have evolved from subjective expert opinion to more data-driven methodologies, such as machine learning[14,15]. Statistical machine learning methods facilitate efficient data exploration in order to identify and analyse disease patterns. These methods are able to encompass the vast array of data generated from birth and patient cohorts, in order to cluster, classify, regress and make predictions from data based on inherent patterns within the large complex dataset. This is in contrast to the traditional methods based on human observation, and testing hypotheses using prior knowledge. Within the context of asthma subtyping, methods such as unsupervised clustering approaches, factor analysis, and principal component analysis have started to be widely used within the last decade. These methods are hypothesis generating, with the overarching notion that the inherent patterns within the data may be a reflection of different underlying aetiologies, genetic basis, and/or immunopathophysiologies, and that identified clusters may represent distinct endotypes of asthma. If this assumption is correct, clustering methodologies could facilitate better understanding of the disease mechanisms, identification of novel therapeutic targets and better clinical trial design incorporating group-specific targeted treatment, all of which are essential steps towards delivery of stratified medicine in asthma.

Here we present a review of the different clustering methodologies - model-free and model-based – and their applications in asthma subtyping. We provide an overview of the major studies and discuss the implications and approaches used.

2.2 What is clustering?

Cluster analysis is a popular unsupervised machine learning method that seeks to identify similar characteristics in subjects (or variables) and group them together on that basis. When selecting groups, the primary aim is to minimize the intra-group variance, while simultaneously maximizing the inter-group variance. Clustering 'classifies' data by labelling objects with cluster 'labels' or giving each object a probability of belonging to a certain cluster. Cluster labels are not known *a priori*, and are derived solely from the data. This is in contrast to supervised methods such as logistic regression and support vector machines, which seek to derive rules for classifying new objects based on a set of already classified objects.

2.2.1 Selection of variables / features and dimension reduction

Cluster analysis lacks the ability to differentiate between clinically relevant and irrelevant variables; thus, choosing the variables to input into the clustering algorithm is one of the most important considerations. Variable or feature selection can be performed subjectively or objectively. Subjective methods choose relevant variables based on expert advice and previous published work. In contrast, objective methods use data-driven approaches to variable/feature selection, the most common of which are stepwise methods (such as backward and forward selection), and dimension reduction techniques (such as principal components analysis [PCA], and factor analysis [FA]). Forward selection progressively adds variables of most significance (based on pre-set p-values) to the model. Backward selection starts with all variables and drops the least significant ones until all the remaining variables are statistically significant.

To reduce the large number of variables, the majority of studies we reviewed employed manual extraction based on expert advice. For example, Moore et al[16] manually reduced 600 variables to 34 by excluding variables with missing data, and those that were either deemed redundant because information was captured by another variable (multicollinearity), or considered to be not clinically relevant. Other studies used dimension reduction techniques such as PCA and FA, which are forms of data reduction that generate small subsets of generally uncorrelated variables from a large dataset of potentially correlated variables. It is useful when we assume there are underlying latent (unobserved) constructs (factors/components) in the data which cannot be measured directly and can influence responses on measured variables. Although these two methods have been used almost interchangeably in the literature we reviewed, there are differences between them. As a general rule of thumb, PCA can be used to reduce data into smaller subsets, while FA can be used to understand what unobserved factors explain the data.

2.2.2 Clustering methods

Three main clustering methods have generally been used in asthma subtyping, including hierarchical approaches, non-hierarchical or partitioning based approaches, and model-based or probabilistic approaches.

Hierarchical clustering

Hierarchical clustering aims to create a pyramidal or (as its name implies) 'hierarchical' grouping of homogeneous clusters that can be displayed in a tree-like graph (dendrogram). It does not require the number of clusters to be specified *a priori*, and cluster assignment is based on similarity of measured characteristics. Within hierarchical clustering there, are two subcategories: agglomerative and divisive methods (Figure 2.1).

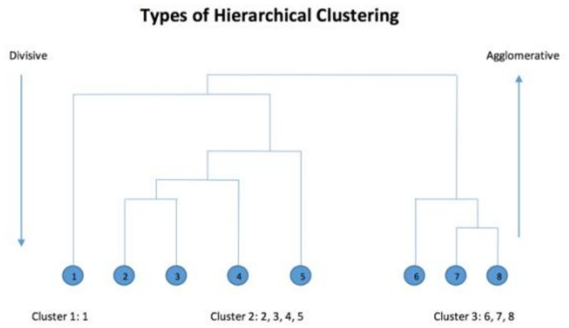


Figure 2.1: *Overview of the difference between agglomerative and divisive hierarchical clustering*

Agglomerative method

The agglomerative method is a bottom-up approach that starts with each data point assigned to its own cluster and iteratively merges the two closest clusters together until all the data belongs to a single cluster [17]. Once clusters are formed, there is no inter cluster switching. The choice of which clusters should be combined is determined by measuring distances, similarities/dissimilarities, and/or using linkage criteria.

This method formulates decisions based on the pattern of variables used without accounting for the overall distribution.

Divisive method

This variant is a top-down approach whereby all objects initially belong to one cluster, which is then recursively divided into sub-clusters until the desired number of clusters is obtained. [18]. By having a single cluster initially, the model gains insight into the spread and type of data and subsequently makes decisions on when and how to divide the sub-clusters.

Similarity/dissimilarity measures

To determine whether objects within the same clustered group are similar or dissimilar, distance measures and linkage criteria (Table 2.1) are used. Distance metrics measure the distance between observations, while linkage criteria measure the distance between clusters. In order to define a similarity measure, the actual similarities between the objects can be evaluated using a distance measure. Choosing a measure for calculating the distance between data can sometimes be arbitrary, as there are no general theoretical guidelines. The Euclidean distance measure, which is the default method in most statistical packages, was used in all studies we reviewed here bar one[19].

Table 2.1: *Most commonly used linkage criteria*

	Linkage Criteria
Centroid	Measures distance between the central point of each cluster
Ward's method	Measures the distance between clusters as the ANOVA sum of squares – i.e. combining information over all cluster members
Complete	Measures the distance between the members of clusters farthest apart

Non-hierarchical clustering

The prototype of non-hierarchical clustering is *k*-means (Figure 2.2). *K*-means is a partitioning method in which the number of clusters is specified a priori and the optimal solution is chosen. It is a variance minimizing algorithm whereby each subject is assigned to its nearest cluster based on the minimal squared Euclidean distance. This method is sensitive to outliers and generally limited to numeric attributes.

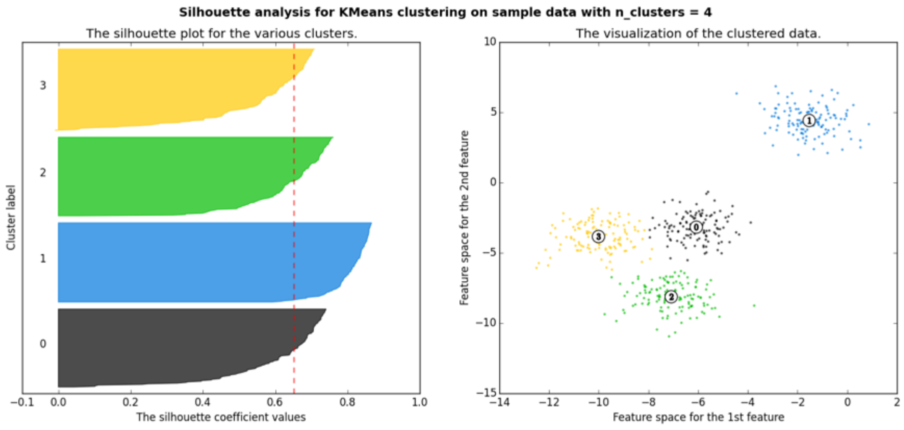


Figure 2.2: *A silhouette plot used for non-hierarchical clustering (k-means) (from[20,21], with permission). A silhouette plot shows how close observations from neighbouring clusters are to each other using a measure of -1 to +1. A value of +1 indicates that that observations may be assigned to the wrong cluster*

Model-based clustering

Model-based clustering (also known as latent class analysis or mixture models), is based on assuming that the observed data are generated by a collection of models, with each cluster corresponding to a different model. Each resulting cluster is represented by a (most commonly) parametric distribution and can be either spherical or ellipsoidal of varying sizes and variance. The advantage of model-based clustering is that it can produce probabilistic cluster assignments for individuals – i.e. it captures the uncertainty in assigning individuals to clusters. Bayesian extensions (e.g. Markov Chain Monte Carlo, expectation-minimization) of model-based clustering can also be used to incorporate prior distributions to reflect uncertainty around model assumptions.

One of the main challenges of model-based clustering is identifying and representing the underlying model assumptions with reasonable complexity. However, unlike a model-free approach, log-likelihood-based statistics such as Bayesian Information Criteria and Model Evidence allow us to select the most parsimonious set of assumptions by penalising model complexity for accuracy. This is in contrast to model-free clustering where an arbitrary distance measure is used to find clusters. Of important note, choosing the best statistically fitting model is not enough; there needs to be an element of expert input into choosing the number of clusters to maximise the potential clinical relevance of the identified subgroups.

2.2.3 Stability of resulting clusters

Cluster stability is an important aspect of validity because cluster methods can generate groups in fairly homogenous data sets. Furthermore, there is always a risk of identifying less meaningful clusters. Stability in this context refers to clusters not disappearing when, for example, outliers are added, data is subset, or random error is introduced to every point to simulate measurement error[22]. The most common way of doing this is to apply the same cluster method to a sample dataset taken from the original one (also termed bootstrapping) and identifying similar clusters using similarity measures.

The similarity values are then compared and stability is taken to be the mean similarity in the new dataset[22].

2.3 Clustering methods in asthma subtyping

2.3.1 The use of principal components analysis/ factor analysis in asthma subtyping

Studies which used PCA/FA as stand-alone analyses for demonstrating the heterogeneity of asthma syndrome and its risk factors are summarised in Table 2.2[23-41]. Sample sizes used in different analyses ranged from 69 to 16,635 and the number of variables used initially ranged from five to 97. The number of resulting components/factors ranged from one to six.

In the context of asthma, the PCA was first used by Smith *et al* to examine whether syndromes of coexisting respiratory symptoms that can be discovered using the response to a large number of questions (>100) from validated questionnaires administered in a birth cohort (Manchester Asthma and Allergy Study - MAAS)[23]. The analysis demonstrated that symptom components (wheeze, cough, wheeze with allergens, wheeze with irritants, chest congestion) were better indicators of the presence and developmental changes in observable secondary asthma phenotypes (such as lung function, airway reactivity and IgE-mediated sensitisation) than the presence of individual symptoms such as wheeze.

Using factor analysis, Bailey *et al*[33] found that intensity of asthma symptoms, asthma management and airflow impairment (FEV_1) were independent components of the disease. This was also seen in the study by Grazzini *et al*[37] where lung function (FEV_1) was an independent factor from asthma symptoms in a mixed teenager-adult population of 69 asthmatics. Lung function was also independent of inflammatory markers (FeNO, sputum eosinophils) in other studies[34,40,41]. The study by Juniper *et al*[38], which included 763 patients older than 12 years who participated in clinical trials, showed that daytime and night-time symptoms despite medication were distinct and independent factors of asthma. Clemmer *et al*[32] used PCA demonstrated that a clinical "endophenotype" relating to corticosteroid responsiveness best predicted corticosteroid response in all replication populations. Other studies in Brazilian[27], British[29], and Japanese[42] children have shown that 'Western diets' were independently associated with an increased risk of wheezing by school age.

Table 2.2*: Studies which used principal components analysis/ factor analysis in asthma subtyping. Avg: average, COREA(Korea): Cohort for Reality and Evolution of Adult Asthma in Korea, COPSAC(Denmark): Copenhagen Prospective Study on Asthma in Childhood, CAMP(US): Childhood Asthma Management Program, IMPACT(US): Improving Asthma Control Trial, PACT(US): Pediatric Asthma Controller Trial, CARE(US): Childhood Asthma Research and Education Network, SOCS(US): Salmeterol or Corticosteroids Study, ACRN(US): Asthma Clinical Research Network, MAAS: Manchester Asthma and Allergy Study*

Cohort/Data setting	Year	Age group	Sample size	# Variables	Method, % variance	Resulting components (PCA)/factors (FA)	Author Reference
Manchester Asthma and Allergy Study	2008	3 5	946 904	21 32	PCA Age 3: 47.5% Age 5: 49.8%	Age 3: 4 Age 5: 5	23
59 rural communities in Ecuador	2011	7-15	Mean 73	29	PCA Component 1: 54.4% Component 2: 50.1% Component 3: 50.7%	2	24
3 Clinical Trials	2012	15-79	1114	21	PCA 76% cumulative	6	25
Generation R study	2012	≤ 4	2173	21	PCA Component 1: 16.3% Component 2: 8.2%	2	26
Education department Sao Francisco do Conde, Brazil	2013	6-12	1307	22	PCA 45.7% cumulative	2	27
COREA	2013	Avg age 70.2 Avg age 44.2	434 1633	11	PCA 53.5% cumulative	Elderly: 4 Non-elderly: 4	28
Manchester Asthma and Allergy Study	2014	Children	1051	97	PCA 15.3% cumulative	3	29
Riyadh Cohort Study	2014	7-17	195	6	PCA 57.3% cumulative	2	30
COPSAC	2015	Neonates	411	5	PCA 41% cumulative	1	31

CAMP, CARE, PACT, ACRN, IMPACT, SOCS	2015	Children	327	6	PCA 100% cumulative	6	[32]
University of Alabama at Birmingham Pulmonary Medicine Clinic	1992	Adults	199	10	FA	3	[33]
Institute of Immunoallergology, Florence IT	1999	16-75	99	8	FA 74.8% cumulative	3	[34]
European Community Respiratory Health Study	2000	20-44	16635	18	FA 58% cumulative	4	[35]
Tucson Children's respiratory Study	2001	6-11	877	25	FA 22.6% cumulative	2	[36]
Stable chronic asthmatics	2001	Adults	69	-	FA 78% cumulative	3	[37]
Salmeterol Quality of Life Study Group	2004	>12	763	21	FA 80.8% cumulative	4	[38]
Health Maintenance Organisation, Kaiser-Permanente US	2005	18-56	2854	53	FA 59% cumulative	5	[39]
Paediatric outpatients, Chinese University of Hong Kong	2005	7-18	92	12	FA 64.6% cumulative	5	[40]
Childhood Asthma Management Program - Clinical trial, Boston USA	2008	5-12	990	17	FA 51.2% cumulative	5	[41]

More recently, both PCA and FA have been used as dimension reduction techniques to generate small subsets from a large number of variables; these small subsets (components/factors) were then used for further clustering. For example, Just *et al* used PCA to reduce 40 variables into 19 that characterised age and BMI, asthma duration, medication use, hospitalisation, atopy, and lung function[43].These were then used in hierarchical clustering. This approach acts as feature extraction in that it can initially visualize/reveal clusters prior to the cluster analysis

2.3.2 Asthma subtype classification with model-free approaches

The studies identified from our literature search which used model-free approaches for subtyping asthma are shown in Table 2.3[16,19,43-62]. Of 22 studies, twelve were carried out in the adult population. Population sample sizes ranged from 57 to 1,843. Method of choice was Ward's hierarchical method with some form of data reduction, whether with PCA, multiple regression analysis, discriminant analysis, factor analysis, or decision trees. K-means clustering was performed in nine out of twenty-two studies, but always as a supplementary method. The resulting number of clusters ranged from two to six.

Paediatric studies

The Trousseau Asthma Program (TAP) in France used Ward's hierarchical clustering as the method of choice[43,51,54]. In the TAP preschool population of 551 wheezers, three clusters of wheezing were identified: mild episodic viral wheeze, atopic multiple-trigger wheeze, and non-atopic uncontrolled wheeze[51]. The mild episodic viral wheeze class was identified in one British[63] and one French cohort[64] using model-based approaches (see below), and the non-atopic uncontrolled wheeze cluster was reproduced in a separate TAP cohort[54]. The multiple-trigger wheeze was previously identified using supervised methods in the Avon Longitudinal Study of Parents and Children (ALSPAC)[65]. This cluster described children with either early or late onset persistent wheezing characterised by atopy and poor lung function. Similar description of wheezing was used in the MAAS cohort to demonstrate that persistent wheezing and multiple early atopy were associated with diminished lung function by age 11 years[66].

Table 2.3*: Studies which used model-free approaches for subtyping asthma. PCA: principal components analysis, FA: factor analysis, Avg: average, SARP: Severe Asthma Research Program (USA), GLAD(UK): GPIAG [General Practitioners in Asthma Group] and Leicester Asthma and Dysfunctional breathing study, TAP: Trousseau Asthma Program (Paris, FR), COREA(Korea): Cohort for Reality and Evolution of Adult Asthma in Korea*

Cohort/Data setting	**Year**	**Age group**	**Sample size, N**	**Data Reduction Technique**	**Method of clustering**	**Number of Clusters**	**Author group reference**
Glenfield Hospital Difficult Asthma Clinic	2008	Avg age: 49.2	184	PCA	2-step: Ward’s hierarchical clustering *k*-means	3	[44]
GLAD		Avg age: 43.4	187			4	
Glenfield Hospital clinical trial		Avg age: 52.4	68			3	
Random selection of patients in Wellington, NZ	2009	25-75	175		2-step: Agglomerative “agnes” clustering Divisive “Diana” Gower’s distance measure Clusters chosen from tree diagram subjectively to include ≥10 subjects per cluster	Agnes: 5 Diana: 4	[45]
SARP	2010	12-80	726		Ward’s hierarchical clustering *Post-hoc* Discriminant analysis for tree-analysis	5	[16]
Asthma Severity Modifying Polymorphisms Project, USA	2010	6-20	154	PCA	2-step: Hierarchical clustering *k*-means clustering	3	[46]

SARP	2011	6-17	161		Ward's hierarchical clustering Centroid linkage *post-hoc* Fisher discriminant analysis- predictors of cluster assignment	4	47
John Hunter Hospital Ambulatory Care Clinic, Newcastle, Australia	2011	19-75	72		Hierarchical clustering Complete linkage	3	48
TAP	2012	6-12	315	PCA	2-step: k-means clustering Ward's hierarchical clustering	3	43
Korean Genome Research Centre for Allergy and Respiratory Diseases cohort	2012	Adults	86		2-step: Hierarchical cluster analysis *k*-means clustering	4	49
NYUBAR, New York City, Bellevue Hospital Center Asthma Clinic	2012	18-75	471		Ward's hierarchical clustering	5	50
TAP	2012	0-3	551		Ward's hierarchical clustering *post-hoc* classification and regression trees Random Forest for predictors of cluster assignment	3	51
TAP	2012	<36 mos	79		Ward's hierarchical clustering	3	52
TALC and BASALT trials, US	2012	Avg age: 37.6	250		Ward's hierarchical clustering *post-hoc* Discriminant analysis for predicting cluster membership	4	53

TAP	2013	5	150		Ward’s hierarchical clustering	4	[54]
COREA	2013	> 18	724		2-step: Ward’s hierarchical clustering	4	[55]
Soonchunhyang University Asthma Genome Research Centre cohort			1843		*k*-means	4	
University of Tsukuba Hospital, Hokkaido University Hospital	2013	16-84	800		Ward’s hierarchical clustering *post-hoc* classification and regression trees Random Forest for predictors of cluster assignment	6	[56]
Quebec City Case Control Asthma Cohort	2013	Avg age: 35.7	377	Factor analysis	2-step: Ward’s hierarchical clustering *k*-means	4	[57]
Niigata University Hospital , Japan	2013	Avg age: 59.8	86	Step-wise multiple regression	Ward’s hierarchical clustering Decision tree analysis for cluster assignment	3	[58]
Paediatric Asthma Clinic, Hacettepe University, Ankara, Turkey	2013	6-18	383	Factor analysis PCA	Hierarchical clustering Gower, Jaccard distances Logistic models	4	[19]
Dutch multicentre study	2013	Adults	200	Factor analysis	Wards hierarchical clustering *k*-means	3	[59]
The Epidemiology and Natural History of Asthma: Outcomes and Treatment Regimens, San Diego US	2014	6-11	518		Ward’s hierarchical clustering	Children: 5	[60]
		>12	3612			Adults: 5	
SARP	2014	Adults	378	INFOGAIN[67] for relevant variables >0.2	*k*-means to partition subjects Ward’s hierarchical clustering	6	[61]

Outpatient clinics, Portugal	2015	Avg age: 45.6	57	Redundant variables removed by Markov Blanket algorithm	Ward’s hierarchical clustering	5 clusters	[62]

The clusters of wheezing described in the TAP cohort remained stable by age 5 years[54]. However, at school age, the clusters were different: 'asthma with severe exacerbations and multiple allergies', 'severe asthma with bronchial obstruction', and 'mild asthma'[43]. These accounted for two "phenotypes": asthma with severe exacerbations, and multiple allergic severe asthma with bronchial obstruction[43]. However, it is important to note that not only were the children from a separate cohort within the TAP, but the clustering methodology was also different; PCA was used for data reduction and a two-step clustering approach including k-means[43]. Furthermore, differing post-hoc analyses were used.

The Severe Asthma Research Program (SARP) is a US multi-centre study comprised of both children and adults with persistent asthma. The study by Fitzpatrick *et al*[47] included 161 children aged 6 to 17 years. Variables were selected subjectively with no data reduction technique, and the authors derived 'composite variables' from binary and questionnaire data discerned by physicians. After Ward's hierarchical clustering, four clusters were identified: 'late-onset symptomatic asthma', 'early onset atopic asthma and normal lung function', 'early onset atopic asthma with mild airflow limitation and comorbidities', 'early onset atopic asthma with advanced airflow limitation'. These results and the accompanying clinical characteristics exhibited by the children were consistent with the previously reported data using clinical observations[68-70]. However, these results differed from findings in a Turkish cohort of children aged 6-18 years with moderate-severe asthma[19]. In contrast to previous studies, the predictive ability of clusters and of original variables in relation to asthma severity in this population was relatively poor[19]. The authors concluded that the search for asthma subtypes needs careful selection of variables which should be consistent across different studies, and that a cautious interpretation of results is warranted[19].

Studies in adults

The initial study that prompted further interest into clustering methodology was done by Haldar *et al* in Leicester, UK[44]. A two-step Ward's hierarchical and subsequent K means cluster analysis was performed in three different datasets (refractory asthmatics from secondary care, primary care data, refractory asthmatics from clinical trial). After variable selection to identify 'most clinically relevant', PCA was performed which reduced the variables into five components. Results of the subsequent cluster analysis revealed three clusters in the primary care dataset and four clusters in the secondary care data. Two clusters were identified in both datasets: 'early onset atopic asthma' and 'obese female with no eosinophilic inflammation'. The primary care dataset identified a third 'benign asthma' cluster, while the secondary care set identified an 'early-onset, symptom-predominant group with minimal eosinophils' cluster as well as a 'late-onset, male predominant, eosinophilic inflammation with few symptoms' cluster. These results were then validated in the clinical trial dataset, which revealed a three-cluster model similar to that in the secondary care set.

Expanding on Haldar's findings, the SARP study[16], which included 726 patients greater than 12 years of age, initially started with 628 variables that were reduced to 34 by excluding missing data, text data, redundant and 'irrelevant' variables. Half of the variables were composite. Ward's method and post-hoc discriminant analysis for Tree analysis was performed to describe five clusters highly determined by frequency of symptoms, medication use, and lung function. Both studies identified a group of obese females with adult onset asthma and less atopy, as well as a group of severe late-onset atopic asthmatics with poor lung function. However, SARP did not use sputum eosinophilia which was an important feature in the study from Leicester. A few years later, the SARP group used a different approach and identified six clusters[61]. K-means clustering partitioned the 378 subjects, while Ward's method clustered the 112 variables into 10 INFOGAIN (measures how well variables predict clusters) ranked variable clusters based on symptoms, atopy, medication use, lung function, corticosteroid use and cause, T_h2inflammation, inflammatory markers, and demographics. Pre-processing of the data included imputing variables with less than 5% missing data while excluding those with more than 5%. Markov blanket algorithms identified redundant variables. Three clusters overlapped with previous results (severe asthmatics, female late-onset with normal lung function) while two were novel (late-onset severe eosinophilic asthmatics with nasal polyps, severe atopic Hispanics). It is interesting to note that similar clusters were seen in children from SARP and Asthma Severity Modifying Polymorphisms[46] project, though the degree of lung function impairment was less.

Patrawalla *et al*[50] based their clustering and variable selection technique on SARP and identified similar clusters to those found by Wu *et al*[61], though the Hispanic females had milder disease. This could be explained by the fact that the sample came from an urban New York City population that had a higher proportion of Hispanics.

The results obtained in Leicester and SARP populations were in part reproduced in a Dutch cohort of severe asthmatic patients that included more thorough inflammatory markers[59]. The resulting three clusters confirmed the existence of two previously reported clusters: severe eosinophilic inflammation-predominant asthma with few symptoms and poor lung function, and obese late onset asthma with low eosinophils additionally provoked by comorbidities such as gastrointestinal oesophageal reflux disease (GORD). The third cluster in the Dutch cohort (mild adult-onset well controlled asthma) which was not found in Leicester and SARP was previously seen in studies in Asian populations which included smoking in their analysis[55,56].

The recurring obesity related subtypes were explored in more detail in two US trials comprising 250 adults[53]. Incorporating detailed data on inflammation, major differences were found between the obese and non-obese populations. Non-obese asthmatics had significantly better lung function. Obese asthmatics with early onset asthma and poor lung function had greater degrees of systemic inflammation (represented by the inverse association between hsCRP and GCRα); this was directly associated with increased glucocorticoid resistance (measured by reduced MKP-1 expression via dexamethasone).

2.3.3 Asthma subtyping and model-based approaches

Latent variable modelling

This topic has recently been reviewed in detail in another review article, which identified a total of 36 studies within the last five years which used model-based approaches to asthma subtyping (four in adult populations, 32 in children)[70]. Sample sizes in these studies ranged from 201 to 11,632. Latent class analysis (14 studies), longitudinal latent class analysis (11 studies), latent class growth analysis (one study), latent growth mixture modelling (8 studies), and mixture models (two studies) methods were used. The number of resulting classes ranged from three to eight, and were in most cases characterised either by physician diagnosed asthma, atopy, and/or fraction of exhaled nitric oxide (FeNO). The most common resulting outcome was “wheeze phenotype”[64,72-83], followed by “atopy class”[64,78,83-88].

Described wheeze classes (often referred to as "phenotypes", although by definition these were not observable, but latent), were either early-onset (transient)[80,89,90] or prolonged[72]), late-onset (characterised as wheeze after age three years persisting into later childhood)[72,76,80,82,85], or persistent (controlled and troublesome), characterised by diminished lung function by school age[9,76]. Early-onset wheeze was found to be predictive of poor lung function, but not atopy, eczema, or rhinitis at age 6-8 years[89]. Late-onset wheeze was associated with bronchial hyperresponsiveness and in some cohorts poorer lung function at age 6 years[65]. The persistent wheeze phenotype was consistently characterized by diminished lung function by school age[9,76].

Atopic sensitisation was the second most common phenotype investigated by latent variable modelling, based on the hypothesis that distinct subtypes may be present. Simpson *et al* applied a hidden Markov chain model to cluster children in MAAS into five different sensitization classes using skin tests and specific IgE data at ages 1, 3, 5, and 8 years[85]. The underlying assumption was that children in each class had the same probability of becoming sensitized or resolving sensitization at each age (and to a similar panel of inhalant and food allergens), and that this differed between classes. Children in one of the four classes (comprising ~25% of sensitised participants), which authors assigned as "multiple early atopy class" were much more likely to have asthma and worse lung function compared to children in all other classes[66,85]. Almost identical five-class model was identified by extending the analysis in MAAS through to age 11 years and in another British birth cohort (Isle of Wight study), indicating stability over time and across different populations [86,91]. However, these classes of sensitisation can be identified only by using statistical inference on longitudinal data, and differentiation between classes at any single cross-sectional point is currently not possible. This emphasises the need to develop diagnostic tools that delineate different classes at any cross-sectional time point among the patient population, to facilitate the applicability of these findings in clinical practice[91-94].

In the adult population, Newby *et al* performed a cluster analysis using mixture models on a multi-centre longitudinal observational study of 349 asthmatics in the British Thoracic Society Severe Refractory Asthma Registry[95]. Variables were initially restricted to those with less than 30% missing data that were non-categorical, and then factor analysis was applied. The resulting five factors (airflow obstruction, exacerbation frequency, IgE/BMI, treatment scaling, blood eosinophilia) were used in the cluster analysis to describe five clusters: 1) 'early onset atopic', 2) 'obese, late onset', 3) 'normal lung function least severe asthma', 4) 'late onset, eosinophilic', 5) 'airflow obstruction'. The best fitting models were chosen by the AIC or BIC, and the clusters were validated using a classifier on a separate dataset from the same registry. Cluster stability for the whole group was only 52%, with cluster two accounting for 71% as the highest, while cluster four accounting for only 25%. A significant proportion of subjects in clusters one, four and five moved to clusters two and three at follow-up indicating greater obesity, lower blood eosinophilia, better lung function, and less exacerbations. Acknowledging small differences in variables used, the results were broadly in accordance with previously reported clusters derived from model-free approaches[16,44]. Gaussian mixture model clustering was also used to investigate cytokine patterns of peripheral blood mononuclear cells' cytokine responses to mite allergens, and the results suggested that asthma is associated with a broad range of immunophenotypes[96]. Various machine learning approaches were also used to identify patterns of IgE responses to a large number of individual allergen molecules in component resolved diagnostics microarrays and associate these with asthma and allergic diseases[14].

2.4 Challenges in asthma clustering

2.4.1 Mixed types of data

Medicine generates many different types of data: binary, numerical, and categorical variables, non-normal distributions, missing values, outliers etc. It is challenging to apply a model that combines these. One possible solution would be to transform the raw variables into one type (i.e. all binary variables). Prosperi *et al*[19] showed that although results were vastly different when comparing the raw and binary variables, they were still clinically consistent with each other. However, it is also important to note that changing continuous variables into binary ones in certain instances would require creating categories. For example, if we take FEV_1 and categorise it based on levels of obstruction (80%, 60-80%, below 60%, above 80%, etc.), we assume that an FEV_1 of 60% has the same clinical significance as an FEV_1 of 79%, which is not necessarily true. Other issues with dichotomizing variables are loss of information leading to a reduction in statistical power, loss of linear relationships between two groups, and underestimation of variability in outcome between groups[97]. Another way to minimize this is to create clinically meaningful categories, but this will likely introduce an element of subjectivity.

2.4.2 Lack of robustness to choice of variables and clustering methods

Different input parameters, even within the same dataset, may produce different results. For example, in the SARP, the same hierarchical clustering techniques on the same dataset produced different clusters[16,47]. The major differences were in the pre-processing of the data and the cluster input. Wu *et al* also included inflammatory markers in their analysis, which would account for better atopy delineation[61].

As mentioned previously, choice of variables has been generally limited to consideration of expert advice based on previous work. Furthermore, there is a practical consideration involved in that the variables chosen have to correspond to the type of data in the cohort given that some studies included all variables[59,61,62] in the dataset while others chose 'the most relevant ones'[43,44,49,55,56,58]. This has resulted in patient exclusion, particularly when there is a requirement to remove variables with missing data. Although some studies implemented imputation techniques in order to overcome this [61,95], the impact on the clinical outcome has not been fully explored, and therefore this should be taken into account when interpreting the results.

In most studies, the choice of distance measure was not specified, thereby assuming the default ones in statistical packages were used (i.e. Euclidean distance). Only two studies[19,45] specified that they varied the distance measures (Gower and/or Jaccard) to observe the effect. One study group used centroid linkage as their similarity measures, whereas the rest were based on Ward. Consequently, we cannot say that the methods employed are the most correct, as there is a repository containing hundreds of options.

Prosperi *et al* hypothesized that resulting clusters from various studies differ due to varying investigator choice of factors, encoding/categorization/transformation of variables, and methodology used[19]. They proceeded to verify this by using different hierarchical clustering and data reduction approaches on a cohort of children aged 6-18 from the Paediatric Asthma Clinic in Ankara, Turkey. Data reduction was done by both FA and PCA, resulting in five 'dimensions' of variables accounting for 35% of the variance. Multiple hierarchical clustering analyses were performed by varying the variable encodings, distance linkages, feature selection, and dimensionality reduction space. Although it was demonstrated that small variations in linkage-distance functions did not affect the resulting clusters[19], they only tested two; it is possible that other linkage criteria could influence the results. What was significant was the fact that changes in variable encodings and transformations resulted in different clusters[19]. It is possible to test the strength of the employed methods by bootstrapping and/or multiple repetitions, however this does not necessarily translate into more plausible overall results.

This is where model-free clustering runs into issues, and whereby a model-based approach might provide more structured methods, as MCMC and EM algorithms are applicable to all modelled distributions. However, in latent class analysis, there is no agreement on the optimal way of determining the number of classes. The most common method is the Bayesian Information Criterion, though other methods such as the Akaike Information Criterion, likelihood tests, bootstrapping, entropy, etc., have been used extensively, thus possibly accounting for the different classes across populations.

2.4.3 Differing subtypes across populations

It is evident that different clusters are identified across different populations (see tables 1 and 2). Other than different statistical methodologies, these disparities may be due to differences in features/variables selected to inform the mode (for example, the choice of lung function variables differed between the studies, and post-bronchodilator FEV_1 was included only in few of these[44]). Of note, as well as influencing heterogeneity in identified clusters, the non-inclusion of some of the potentially important variables (e.g. post-bronchodilator lung function) may result in a failure to capture some important underlying mechanisms. Additionally, most studies were done in severe or moderate-severe asthmatics, and the same subtypes may not be seen in the mild asthma population.

It is also important to note that clusters identified cross-sectionally at a specific time point may not always be seen at different time points. Further longitudinal analysis is required to visualize how the clusters vary over time.

2.5 Conclusion

The understanding of asthma has come a long way, and data-driven hypothesis-generating clustering methods have aided in identifying distinct subtypes of asthma. However, we must be careful when translating these results into clinical practice, as one needs to use statistical inference on large data set to identify disease subtypes, and biomarkers that would allow differentiation of such subtypes at any cross-sectional time point are in most cases not available. Further challenges to the optimal use of clustering methodologies include tailoring models to individual datasets and incorporating genetic, epigenetic and more detailed molecular level data. Resulting models should then be able to accommodate large volumes of data in order to discern the developmental profiles of each individual, facilitating genuine personalised approach to asthma management.

2.6 References

1. *From the Global Strategy for Asthma Management and Prevention (GINA 2015), URL: http://www.ginasthma.org/*. 2015.
2. Wenzel, S.E., *Asthma: defining of the persistent adult phenotypes.* Lancet, 2006. **368**(9537): p. 804-13.
3. Anderson, G.P., *Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease.* Lancet, 2008. **372**(9643): p. 1107-1119.
4. Lotvall, J., et al., *Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome.* J Allergy Clin Immunol, 2011. **127**(2): p. 355-60.
5. Rackemann, F.M., *A clinical survey of 1074 patients with asthma followed for two years.* The Journal of Laboratory and Clinical Medicine, 1927: p. 1185-1197.
6. Custovic, A., et al., *Exercise testing revisited. The response to exercise in normal and atopic children.* Chest, 1994. **105**(4): p. 1127-32.
7. Vonk, J.M., et al., *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): p. 322-7.
8. Pavord, I.D. and A. Agusti, *Blood eosinophil count: a biomarker of an important treatable trait in patients with airway disease.* Eur Respir J, 2016 in press.
9. Belgrave, D.C., A. Custovic, and A. Simpson, *Characterizing wheeze phenotypes to identify endotypes of childhood asthma, and the implications for future management.* Expert Rev Clin Immunol, 2013. **9**(10): p. 921-36.
10. Martinez, F.D., et al., *Asthma and wheezing in the first six years of life. The Group Health Medical Associates.* N Engl J Med, 1995. **332**(3): p. 133-8.
11. Lowe, L.A., et al., *Wheeze phenotypes and lung function in preschool children.* Am J Respir Crit Care Med, 2005. **171**(3): p. 231-7.
12. Wenzel, S.E., *Asthma phenotypes: the evolution from clinical to molecular approaches.* Nat Med, 2012. **18**(5): p. 716-25.
13. Custovic, A., et al., *The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together.* Thorax, 2015.
14. Prosperi, M.C., et al., *Challenges in interpreting allergen microarrays in relation to clinical symptoms: a machine learning approach.* Pediatr Allergy Immunol, 2014. **25**(1): p. 71-9.
15. Prosperi, M.C., et al., *Predicting phenotypes of asthma and eczema with machine learning.* BMC Med Genomics, 2014. **7 Suppl 1**: p. S7.
16. Moore, W.C., et al., *Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program.* Am J Respir Crit Care Med, 2010. **181**(4): p. 315-23.
17. Duda, R., Hart P, *Pattern Classification and Scene Analysis*. 1973: Wiley.
18. Rokach, L., Oded M, *Clustering Methods*, in *Data Mining and Knowledge Discovery Handbook*. 2005, Springer. p. 321-352.
19. Prosperi, M.C., et al., *Challenges in identifying asthma subgroups using unsupervised statistical learning techniques.* Am J Respir Crit Care Med, 2013. **188**(11): p. 1303-12.
20. Pedregosa, e.a., *Scikit-Learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.
21. al, P.e., *Scikit-Learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.

22. C, H., *Cluster-wise assessment of cluster stability*. 2006, University College London: London UK.
23. Smith, J.A., et al., *Dimensions of respiratory symptoms in preschool children: population-based birth cohort study.* Am J Respir Crit Care Med, 2008. **177**(12): p. 1358-63.
24. Rodriguez, A., et al., *Urbanisation is associated with prevalence of childhood asthma in diverse, small rural communities in Ecuador.* Thorax, 2011. **66**(12): p. 1043-50.
25. Greenberg, S., et al., *Airway obstruction lability helps distinguish levels of disease activity in asthma.* Respir Med, 2012. **106**(4): p. 500-7.
26. Tromp, II, et al., *Dietary patterns and respiratory symptoms in pre-school children: the Generation R Study.* Eur Respir J, 2012. **40**(3): p. 681-9.
27. de Cassia Ribeiro Silva, R., et al., *Dietary Patterns and Wheezing in the Midst of Nutritional Transition: A Study in Brazil.* Pediatr Allergy Immunol Pulmonol, 2013. **26**(1): p. 18-24.
28. Park, H.W., et al., *Differences between asthma in young and elderly: results from the COREA study.* Respir Med, 2013. **107**(10): p. 1509-14.
29. Patel, S., et al., *Cross-sectional association of dietary patterns with asthma and atopic sensitization in childhood - in a cohort study.* Pediatr Allergy Immunol, 2014. **25**(6): p. 565-71.
30. Al-Daghri, N.M., et al., *Th1/Th2 cytokine pattern in Arab children with severe asthma.* Int J Clin Exp Med, 2014. **7**(8): p. 2286-91.
31. Chawes, B.L., et al., *Neonates with reduced neonatal lung function have systemic low-grade inflammation.* J Allergy Clin Immunol, 2015. **135**(6): p. 1450-6 e1.
32. Clemmer, G.L., et al., *Measuring the corticosteroid responsiveness endophenotype in asthmatic patients.* J Allergy Clin Immunol, 2015. **136**(2): p. 274-81 e8.
33. Bailey, W.C., et al., *Asthma severity: a factor analytic investigation.* Am J Med, 1992. **93**(3): p. 263-9.
34. Rosi, E., et al., *Sputum analysis, bronchial hyperresponsiveness, and airway function in asthma: results of a factor analysis.* J Allergy Clin Immunol, 1999. **103**(2 Pt 1): p. 232-7.
35. Sunyer, J., et al., *International assessment of the internal consistency of respiratory symptoms. European Community Respiratory Health Study (ECRHS).* Am J Respir Crit Care Med, 2000. **162**(3 Pt 1): p. 930-5.
36. Holberg, C.J., et al., *Factor analysis of asthma and atopy traits shows 2 major components, one of which is linked to markers on chromosome 5q.* J Allergy Clin Immunol, 2001. **108**(5): p. 772-80.
37. Grazzini, M., et al., *Relevance of dyspnoea and respiratory function measurements in monitoring of asthma: a factor analysis.* Respir Med, 2001. **95**(4): p. 246-50.
38. Juniper, E.F., et al., *Relationship between quality of life and clinical status in asthma: a factor analysis.* Eur Respir J, 2004. **23**(2): p. 287-91.
39. Schatz, M., et al., *Relationships among quality of life, severity, and control measures in asthma: an evaluation using factor analysis.* J Allergy Clin Immunol, 2005. **115**(5): p. 1049-55.
40. Leung, T.F., et al., *Clinical and atopic parameters and airway inflammatory markers in childhood asthma: a factor analysis.* Thorax, 2005. **60**(10): p. 822-6.
41. Holt, E.W., et al., *Identifying the components of asthma health status in children with mild to moderate asthma.* J Allergy Clin Immunol, 2008. **121**(5): p. 1175-80.

42. Miyake, Y., et al., *Maternal dietary patterns during pregnancy and risk of wheeze and eczema in Japanese infants aged 16-24 months: the Osaka Maternal and Child Health Study.* Pediatr Allergy Immunol, 2011. **22**(7): p. 734-41.
43. Just, J., et al., *Two novel, severe asthma phenotypes identified during childhood using a clustering approach.* Eur Respir J, 2012. **40**(1): p. 55-60.
44. Haldar, P., et al., *Cluster analysis and clinical asthma phenotypes.* Am J Respir Crit Care Med, 2008. **178**(3): p. 218-24.
45. Weatherall, M., et al., *Distinct clinical phenotypes of airways disease defined by cluster analysis.* Eur Respir J, 2009. **34**(4): p. 812-8.
46. Benton, A.S., et al., *Overcoming heterogeneity in pediatric asthma: tobacco smoke and asthma characteristics within phenotypic clusters in an African American cohort.* J Asthma, 2010. **47**(7): p. 728-34.
47. Fitzpatrick, A.M., et al., *Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program.* J Allergy Clin Immunol, 2011. **127**(2): p. 382-389 e1-13.
48. Baines, K.J., et al., *Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples.* J Allergy Clin Immunol, 2011. **127**(1): p. 153-60, 160 e1-9.
49. Jang, A.S., et al., *Identification of subtypes of refractory asthma in Korean patients by cluster analysis.* Lung, 2013. **191**(1): p. 87-93.
50. Patrawalla, P., et al., *Application of the asthma phenotype algorithm from the Severe Asthma Research Program to an urban population.* PLoS One, 2012. **7**(9): p. e44540.
51. Just, J., et al., *Novel severe wheezy young children phenotypes: boys atopic multiple-trigger and girls nonatopic uncontrolled wheeze.* J Allergy Clin Immunol, 2012. **130**(1): p. 103-10 e8.
52. Gouvis-Echraghi, R., et al., *Exhaled nitric oxide measurement confirms 2 severe wheeze phenotypes in young children from the Trousseau Asthma Program.* J Allergy Clin Immunol, 2012. **130**(4): p. 1005-7 e1.
53. Sutherland, E.R., et al., *Cluster analysis of obesity and asthma phenotypes.* PLoS One, 2012. **7**(5): p. e36631.
54. Just, J., et al., *Wheeze phenotypes in young children have different courses during the preschool period.* Ann Allergy Asthma Immunol, 2013. **111**(4): p. 256-261 e1.
55. Kim, T.B., et al., *Identification of asthma clusters in two independent Korean adult asthma cohorts.* Eur Respir J, 2013. **41**(6): p. 1308-14.
56. Kaneko, Y., et al., *Asthma phenotypes in Japanese adults - their associations with the CCL5 and ADRB2 genotypes.* Allergol Int, 2013. **62**(1): p. 113-21.
57. Lavoie-Charland, m., et al., *Multivariate Asthma Phenotypes in Adults: The Quebec City Case-Control Asthma Cohort.* Open Journal of Respiratory Diseases, 2013. **Vol.03No.04**: p. 10.
58. Sakagami, T., et al., *Cluster analysis identifies characteristic phenotypes of asthma with accelerated lung function decline.* J Asthma, 2014. **51**(2): p. 113-8.
59. Amelink, M., et al., *Three phenotypes of adult-onset asthma.* Allergy, 2013. **68**(5): p. 674-80.
60. Schatz, M., et al., *Phenotypes determined by cluster analysis in severe or difficult-to-treat asthma.* J Allergy Clin Immunol, 2014. **133**(6): p. 1549-56.

61. Wu, W., et al., *Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data.* J Allergy Clin Immunol, 2014. **133**(5): p. 1280-8.
62. Loureiro, C.C., et al., *Cluster analysis in phenotyping a Portuguese population.* Rev Port Pneumol (2006), 2015.
63. Spycher, B.D., et al., *Distinguishing phenotypes of childhood wheeze and cough using latent class analysis.* Eur Respir J, 2008. **31**(5): p. 974-81.
64. Herr, M., et al., *Risk factors and characteristics of respiratory and allergic phenotypes in early childhood.* J Allergy Clin Immunol, 2012. **130**(2): p. 389-96 e4.
65. Henderson, J., 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): p. 974-80.
66. Belgrave, D.C.M., et al., *Trajectories of lung function during childhood.* American journal of respiratory and critical care medicine, 2014. **189**(9): p. 1101-9.
67. Bossley, C.J., et al., *Corticosteroid responsiveness and clinical characteristics in childhood difficult asthma.* Eur Respir J, 2009. **34**(5): p. 1052-9.
68. Chipps, B.E., et al., *Demographic and clinical characteristics of children and adolescents with severe or difficult-to-treat asthma.* J Allergy Clin Immunol, 2007. **119**(5): p. 1156-63.
69. Bacharier, L.B., et al., *Classifying asthma severity in children: mismatch between symptoms, medication use, and lung function.* Am J Respir Crit Care Med, 2004. **170**(4): p. 426-32.
70. Howard, R., et al., *Distinguishing Asthma Phenotypes Using Machine Learning Approaches.* Curr Allergy Asthma Rep, 2015. **15**(7): p. 38.
71. Savenije, O.E., et al., *Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA.* J Allergy Clin Immunol, 2011. **127**(6): p. 1505-12 e14.
72. Chen, Q.e.a., *Using latent class growth analysis to identify childhood wheeze phenotypes in an urban birth cohort.* Annals of Allergy, Asthma, and Immunology, 2012. **108**(5): p. 311-315.
73. Weinmayr, G., et al., *Asthma phenotypes identified by latent class analysis in the ISAAC phase II Spain study.* Clin Exp Allergy, 2013. **43**(2): p. 223-32.
74. Spycher, B.D., et al., *Comparison of phenotypes of childhood wheeze and cough in 2 independent cohorts.* J Allergy Clin Immunol, 2013. **132**(5): p. 1058-67.
75. Belgrave, D.C., et al., *Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing.* J Allergy Clin Immunol, 2013. **132**(3): p. 575-583.e12.
76. Cano-Garcinuno, A., I. Mora-Gandarillas, and S.S. Group, *Wheezing phenotypes in young children: an historical cohort study.* Prim Care Respir J, 2014. **23**(1): p. 60-6.
77. Panico, L., et al., *Asthma trajectories in early childhood: identifying modifiable factors.* PLoS One, 2014. **9**(11): p. e111922.
78. Depner, M., et al., *Clinical and epidemiologic phenotypes of childhood asthma.* Am J Respir Crit Care Med, 2014. **189**(2): p. 129-38.
79. Caudri, D., et al., *Perinatal risk factors for wheezing phenotypes in the first 8 years of life.* Clin Exp Allergy, 2013. **43**(12): p. 1395-405.
80. Lodge, C.J., et al., *Childhood wheeze phenotypes show less than expected growth in FEV1 across adolescence.* Am J Respir Crit Care Med, 2014. **189**(11): p. 1351-8.

81. Savenije, O.E., et al., *Association of IL33-IL-1 receptor-like 1 (IL1RL1) pathway polymorphisms with wheezing phenotypes and asthma in childhood.* J Allergy Clin Immunol, 2014. **134**(1): p. 170-7.

82. Belgrave, D.C., et al., *Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies.* PLoS Med, 2014. **11**(10): p. e1001748.

83. Siroux, V., et al., *Identifying adult asthma phenotypes using a clustering approach.* Eur Respir J, 2011. **38**(2): p. 310-7.

84. Simpson, A., et al., *Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study.* Am J Respir Crit Care Med, 2010. **181**(11): p. 1200-6.

85. Lazic, N., et al., *Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts.* Allergy, 2013. **68**(6): p. 764-70.

86. Garden FL, S.J., Marks G, *Atopu phenotypes in he Childhood Asthma Prevention Study (CAPS) cohort and the relationship with allergic disease: clinical mechanisms in allergic disease.* Journal of the Biritish Society of Allergy and Clinical Immunology, 2013. **43**(6): p. 633-641.

87. Havstad, S., et al., *Atopic phenotypes identified with latent class analyses at age 2 years.* J Allergy Clin Immunol, 2014. **134**(3): p. 722-727 e2.

88. Savenije, O.E., et al., *Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA.* The Journal of allergy and clinical immunology, 2011. **127**(6): p. 1505-12.e14.

89. Savenije, O.E., et al., *Association of IL33-IL-1 receptor-like 1 (IL1RL1) pathway polymorphisms with wheezing phenotypes and asthma in childhood.* Journal of Allergy and Clinical Immunology, 2014.

90. Custovic, A., N. Lazic, and A. Simpson, *Pediatric asthma and development of atopy.* Curr Opin Allergy Clin Immunol, 2013. **13**(2): p. 173-80.

91. Holt, P.G., et al., *Distinguishing benign from pathologic TH2 immunity in atopic children.* J Allergy Clin Immunol, 2016. **137**(2): p. 379-87.

92. Custovic, A., et al., *Evolution pathways of IgE responses to grass and mite allergens throughout childhood.* J Allergy Clin Immunol, 2015. **136**(6): p. 1645-52 e1-8.

93. Simpson, A., et al., *Patterns of IgE responses to multiple allergen components and clinical symptoms at age 11 years.* J Allergy Clin Immunol, 2015. **136**(5): p. 1224-31.

94. Newby, C., et al., *Statistical cluster analysis of the British Thoracic Society Severe refractory Asthma Registry: clinical outcomes and phenotype stability.* PLoS One, 2014. **9**(7): p. e102987.

95. Wu, J., et al., *Relationship between cytokine expression patterns and clinical outcomes: two population-based birth cohorts.* Clin Exp Allergy, 2015. **45**(12): p. 1801-11.

96. Altman, D.G. and P. Royston, *The cost of dichotomising continuous variables.* BMJ, 2006. **332**(7549): p. 1080.

97. Mitchell, T., *Machine Learning*. 1997: McGraw-Hill Science.

Chapter 3 Asthma phenotypes in childhood

Matea Deliu, Danielle Belgrave, Matthew Sperrin, Iain Buchan, Adnan Custovic
Expert Rev Clin Immunol. 2017 Jul;13(7):705-713. doi: 10.1080/1744666X.2017.1257940.

3.1 Abstract

Introduction: Asthma is no longer thought of as a single disease, but rather a collection of varying symptoms expressing different disease patterns. One of the ongoing challenges is understanding the underlying pathophysiological mechanisms that may be responsible for the varying responses to treatment.

Areas Covered: This review provides an overview of our current understanding of the asthma phenotype concept in childhood and describes key findings from both conventional and data-driven methods.

Expert Commentary: With the vast amounts of data generated from cohorts, there is hope that we can elucidate distinct pathophysiological mechanisms, or endotypes. In return, this would lead to better patient stratification and disease management, thereby providing true personalised medicine.

3.2 Introduction

Asthma is a common disease which has been rapidly increasing in both prevalence and incidence. The surge of new cases, particularly in the western world, can be in part attributed to rapid changes in lifestyle and environmental exposures. However, the exact extent of this environmental impact is yet to be fully understood. For the most part, asthma starts in early childhood, though some patients can develop more severe disease symptoms either in teenage years or adulthood; the incidence of new cases is lower in adults[1,2]. The variability in the expression of asthma symptoms observed within a clinical setting has prompted the move away from the concept that asthma is a single entity; it is now generally considered that asthma is an umbrella term for several distinct conditions that share common clinical features such as wheezing, cough, shortness of breath, and variable airflow obstruction[3].

A phenotype is defined as an observable property or trait that arises from the interaction of genes and environmental exposures. Phenotypes are therefore characteristics that can be directly observed and measured (either biochemically or physically)[4]. For example, in clinical terms 'trait' may refer to wheeze or lung function; wheeze can be auscultated and lung function can be measured, e.g. by spirometry. However, these traits can vary drastically in terms of the manner in which they are manifested between different patients, and relating them to underlying mechanisms would be essential to understand the pathology of the disease(s). With such variation in the clinical expression of asthma, the concept of 'endotype' has been proposed[5,6]. Whereas the term 'phenotype' refers to an external, directly observable characteristic, an 'endotype' indicates a subtype of the disease with a distinct underlying pathophysiologic mechanism, which may in part explain the observed heterogeneity in phenotype manifestation[7]. As shown in Figure 1, our task in understanding asthma is to disentangle the multifarious phenotypes in an attempt to distinguish distinct subtypes which may in turn indicate the presence of distinct endotypes which have distinct causal mechanisms. Multiple 'endotypes' can therefore give rise to the same or similar phenotype[7], and this endotype-phenotype connection itself may not be a static characteristic. A level of complexity is added by the fact that some of the key phenotypes (e.g. eosinophilic inflammation) are highly variable in childhood[8].

The importance of identifying asthma endotypes (or subtype) is primarily for understanding disease mechanisms. Our current treatment guidelines are still guided primarily by symptoms and lung function, yet we have all seen that similar symptoms have different levels of response to commonly used treatments. Deciphering the cause of this heterogeneity would lead to better treatment targeting, and would be a step towards "precision" medicine. Furthermore, there is hope that this knowledge would pave the way for better predictive modelling based on risk factors and disease progression. In other words, if we can correctly categorise children into different subtypes at a young age, we will have insight into the developmental trajectory of the disease in order to apply preventative strategies. However, we have yet to fully discover all the modifiable factors that influence the natural course of the asthma-related diseases.

A wide array of methodologies have been utilised to enhance subtypying of asthma and allergic diseases in childhood. Such methods range from expert investigator-led approaches to identifying patterns and co-occurrence of symptoms which mimics the approach used within a clinical setting, to supervised statistical modelling approaches which aim to test hypothesised frameworks for disease profiles, and unsupervised data-driven statistical methods which take an agnostic view of the data in order to infer structure based on pattern recognition computer algorithms[9,10]. Combinations of supervised and unsupervised models allow us to incorporate prior clinical knowledge within a data-driven approach and can also enable us to evaluate the likelihood that our observed data is aligned to prior hypotheses or clinical assumptions. All of these methods have advanced our knowledge in this field in different ways. Subjective expert-driven approaches have been able to describe and externally validate what is seen in clinic, while data driven machine learning approaches are capitalizing on the wealth of data available by seeking to find patterns of disease and then applying that to the general population.

Our understanding of asthma heterogeneity is constantly evolving. In this review, we present our current knowledge of the subtypes in childhood asthma. We explain the clinical implications of phenotyping and subtyping asthma in childhood, as this is the critical period that is amenable to potential prevention strategies.

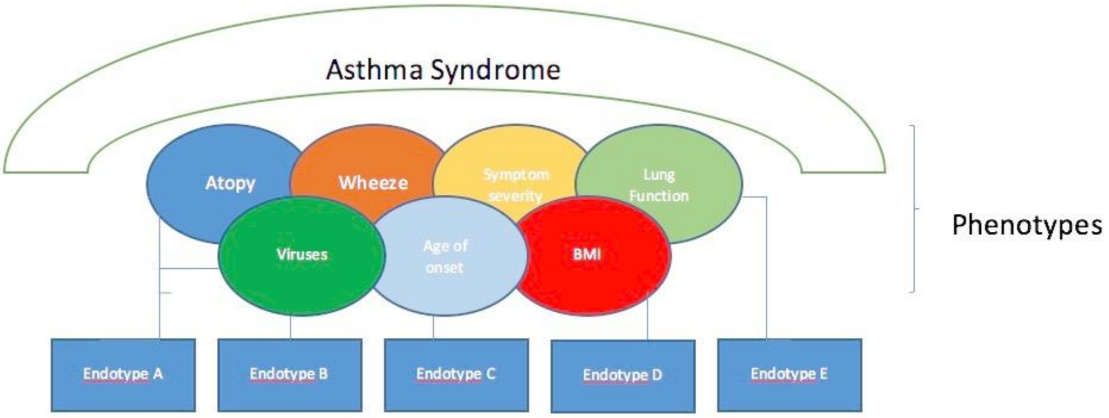


***Figure 3.1**: Schematic drawing of the asthma syndrome. BMI: body mass index*

3.3 Wheeze phenotypes

The presence or absence of wheeze is one of the key determinants of asthma. Wheezing is associated with airflow obstruction due to airway narrowing. In children, this auscultatory finding is primarily an expression of large airway obstruction. Wheeze can manifest in various ways and at different time points in a child's life. However, it should be noted that not all wheeze can lead to asthma, particularly in those children below the age of 3 years. Within the last decade, there has been a surge of data describing various ways of identifying wheeze phenotypes in childhood. At the most general level, phenotypes have been characterised according to age of onset (early vs late), according to severity, or based on triggers (viral, allergen, or other).

3.3.1 Wheeze phenotypes based on age of onset and remission (temporal pattern)

Differentiating wheeze phenotypes by age of onset during childhood can give insight into the pattern of disease expression later in adulthood. The ability to predict both the timing (whether a child will develop wheezing early on or later in life) as well as subsequent profile of symptom development may allow us to pre-emptively manage such occurrence and severity. As a result, age of onset of wheeze and its persistence have become key determinants for identifying distinct wheeze subtypes.

Early onset transient wheeze

The first seminal paper to characterize wheeze and its natural history was by Martinez et al[11] in an unselected birth cohort based in Tucson, AZ. Using a subjective clinical approach based on the observed patterns of wheeze over time, 'transient-early onset wheeze' was defined as wheeze with onset before the age of three years, with subsequent resolution by age six years[11]. The authors hypothesized that transient early wheeze may be triggered by viral infections. These children initially had poorer lung function in infancy which later improved, although remaining lower compared to the control group.

The Manchester Asthma and Allergy Study (MAAS) used a longitudinal latent class model integrating data from both parental questionnaires and medical records to ascertain children's wheeze symptoms. This model represented a statistical data-driven approach which assigns children to their most probable latent (unobserved) cluster based on the patterns of wheeze at multiple time-points (Figure 3.2). These clusters are hypothesised to represent distinct symptom profiles with distinct underlying pathophysiology. Using complementary data sources enabled the modelling of uncertainty in parental or physician diagnosis of wheeze. Similar to the Tuscon cohort, MAAS also identified a transient early wheeze group (wheezing from age 1-5). However, unlike Tuscon, they found that lung function in these children remained impaired compared to non-wheezers throughout childhood[12].

Using a similar statistical approach, but based only on parental reporting of wheeze, Henderson et al[13] of the Avon Longitudinal Study of Parents and Children (ALSPAC) (see Figure 3.2) cohort identified a prolonged early wheezing group (wheezing from 6-54 months). Compared to the findings from the Tucson study, the prolonged-early phenotype is thought to be a more severe form of transient-early wheeze based on observed diminished lung function at age 8-9 years for this group of children. However early-life lung function was not available thus hindering evaluation as to whether there is a distinction in pre-morbid lung function for this group of children compared to the transient early wheezers. Indeed, as lung function measurements are difficult to obtain below the age of 2 years, many institutions are increasingly reluctant on labelling early wheeze as asthma. The results from the ALSPAC have been replicated using similar methodology in other cohorts [12,14-17]. Within the various studies described, the definition and prevalence of early-onset wheeze varied between study groups (in part likely a consequence of the age at which data was collected). For example, the prevalence of early-onset wheeze in the Tucson study[11] was 19.9%, and 23.5% in MAAS[18]. Using a solely data-driven approach, the proportion of children in MAAS[19] rose to 29.3%, while ALSPAC reported 42.4%[13].

In most studies, the majority of children who wheezed early in infancy have symptom resolution in later childhood [11,20-23]. Risk factors identified for transient early wheeze include exposure to tobacco smoke, day care attendance, virus infections, and family history of asthma[14,24,25]. Results have been inconsistent with regards to sex and position in sibship[25-27].

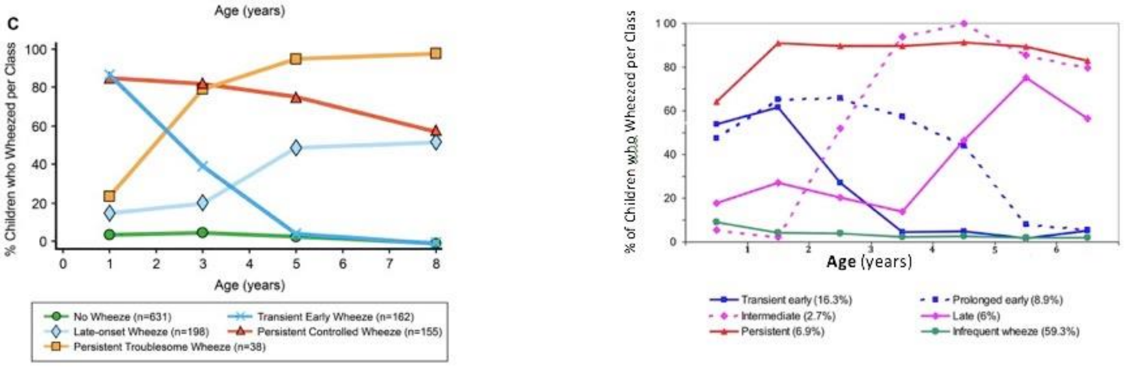


***Figure 3.2**: Comparison of wheeze phenotypes from Manchester Asthma and Allergy Study (MAAS) and Avon Longitudinal Study of Parents and Children (ALSPAC) (adapted and modified from [12] and [28], with permission).*

Persistent wheeze

Similar to transient wheezers, children with persistent wheeze start wheezing in early life; however, in contrast to transient wheezers, their symptoms do not resolve, but continue in later childhood[11]. One of the characteristics of persistent wheeze is diminished lung function by school age, with likely persistence to adulthood. Martinez et al[11] found that compared to controls (non-wheezers), children with persistent wheeze initially had normal lung function, but that lung function significantly worsened by age six years. This was also seen at age 11 years in a follow-up study[29]. These children were more likely to be atopic, have higher IgE levels, and be sensitized early on. Using similar subjective methods to characterise children, study groups from New Zealand[20] and Australia[30], observed that children who were categorised as persistent wheezers continued to wheeze into adulthood and had consistently lower lung function.

Availability of data from primary care medical records in the MAAS cohort[31] allowed stratification of persistent wheeze into two distinct subgroups: persistent controlled and persistent troublesome wheeze. Children with "troublesome" wheeze were more likely to have a high symptom burden despite high doses of inhaled corticosteroids. Belgrave *et al*[12] found that being highly atopic (whether to food or inhalant allergens) and having concurrent eczema were strong predictors of this phenotype. The persistent wheeze identified in ALSPAC showed a similar pattern though with weaker atopy associations[13]. Early identification of children who are at risk of persistent troublesome wheeze would allow for better management, and possible prevention to reduce the likelihood of persistence of troublesome symptoms into adulthood.

An intermediate onset wheeze (onset between 18-42 months) was identified in ALSPAC, though given the time of onset, this could be classified as early in other studies. Children in this category were more likely to be atopic (particularly to house dust mite and cat), have poor lung function, and subsequently be at risk of developing asthma in later childhood.

Atopic sensitization is a well-documented risk factor for both persistent wheezing[11-13] and the persistence of asthma from school-age to late teenage years[32]. However, Simpson et al have demonstrated that atopic sensitisation may not be a simple dichotomous trait, but rather a collection of several different atopic vulnerabilities[33]. Of the four distinct sensitization classes identified in this study, there was a strong association between only one of these classes (assigned as multiple early atopy) and persistent wheezing. Similar structure within the data on sensitisation, and identical association with persistent wheezing and asthma has been reported in the Isle of Wight study[34]. Using a similar approach, similar atopic patterns were identified in the Childhood Asthma Prevention Study (CAPS) from Australia, and asthma at age 8 was associated with mixed food (predominantly peanut) and inhalant sensitization[35]. However, such stratification of atopic sensitization requires the use of complex machine learning models on rich longitudinal data, and cross-sectional biomarkers of different atopic vulnerabilities will need to be discovered if this is to be translated into clinical practice.

Late onset wheeze

Late onset wheeze is generally described as wheeze with onset after age three years which persists into later childhood. Atopic sensitization is consistently associated with this phenotype of wheeze across the different studies[12-14]. In ALSPAC, this sensitisation was grass pollen induced[13]. However, the association between late onset wheezing, lung function and bronchial hyperresponsiveness has differed between different studies. For example, MAAS and ALSPAC found that late onset wheeze was significantly associated with increased bronchial hyperresponsiveness and lung function impairment at age 6[13,31] while Prevention and Incidence of Asthma and Mite Allergy (PIAMA)[15] and Southampton Women's Study (SWS)[14] showed such association. Amongst environmental exposures, maternal smoking during pregnancy was risk factor for late onset wheezing in some[11,36], but not all studies[12,14].

Little is known on the stability of this phenotype, and further longitudinal follow-up is required to determine whether late-onset wheeze persists into adulthood.

3.3.2 Wheeze phenotypes based on triggers

Treatment of early childhood wheeze has been generally limited by the lack of evidence for the efficacy of most currently available treatments. To facilitate management of young children with wheezing, the European Task Force[37] proposed a clinical differentiation of early childhood wheezing into two subgroups: ‘episodic viral wheeze (EVW)’ and ‘multiple trigger wheeze (MTW)’. EVW is described as intermittent seasonal wheeze episodes with occasional periods of ‘feeling well’. Children labelled with episodic viral wheeze tend to be non-atopic with almost normal lung function, and symptoms tend to resolve by late childhood[38,39]. Multiple trigger wheeze (MTW) is defined as “wheezing that shows discreet exacerbations, but also symptoms in between episodes[40].” Proposed triggers of MTW includes allergens, exercise, mist, crying, laughter etc.[37]. Of note, the major trigger of MTW are respiratory virus infections, making a clear distinction between EVW and MTW difficult. Both EVW and MTW are defined cross-sectionally at different ages, with severity of symptoms being accounted for by varying frequencies of wheezing episodes (more than three classified as frequent)[41,42]. This classification was based on expert opinion, rather than solid data, and the implications were that ICS treatment would be more appropriate and effective for the MTW compared to EVW. However, neither of these two proposed “phenotypes” have been shown to be stable. For example, Schultz et al[43] showed that more than 50% of children change their phenotypic classification within 12 months of allocation into one of these two groups; EVW most frequently changed to MTW. However, a few years later, Schultz and Brand showed that regular use of corticosteroids had a modest effect on reducing symptoms in the EVW group[44]. Despite this, it is important to note that a phenotype-driven approach to treatment is still currently limited by our ability to accurately differentiate phenotypes and therefore the clinical utility is yet to be determined[44].

There have been attempts to validate this classification using data-driven techniques. For example, EVW was identified using latent class analysis in two studies from the Trousseau Asthma Program (TAP)[38,45], while one study from Leicester, UK[39] suggested that the 'transient early wheeze' phenotype was very similar to EVW. A more severe type of EVW characterised by high FeNO levels and strong family history of asthma was described by Kappelle et al[46], and these children were at an increased risk of developing asthma at age five to 10 years, which was found to persist into adulthood[47]. The TAP group identified an atopic MTW in boys that was associated with severe wheezing and allergic comorbidities[45].

There have been attempts to reconcile phenotypes described through temporal patterns of symptoms and those based on triggers. For example, it has been suggested that transient early wheeze may correspond to EVW, and persistent wheeze to MTW[11,39]. However, identifying children as having EVW/MTW is a relatively poor predictor of whether they would subsequently be classified as transient/persistent wheezers, with low positive predictive values[41,48-50]. Depner and colleagues[16] sought to compare the data-driven wheeze phenotypes (early, transient, intermediate, late-onset, persistent) with the clinical classification of EVW/MTW by using a multi-country cohort of children from birth to 6 years of age and applying a longitudinal latent class model. Their results showed that approximately 60% of children with EVW were found to be within the transient early wheeze class implying that this phenotype is unlikely to encompass a chronic condition with poor lung function, but rather an initial response to viruses that eventually resolves[16]. However, the correlation with the LCA classes was poor. In contrast, 60-70% of children with MTW were either in the intermediate, late-onset, or persistent wheeze class. The correlation with the LCA classes was very high. Further to this, a variant of MTW was also identified and labelled recurrent unremitting wheeze (symptoms or wheeze without a cold on multiple occasions). This was characterised by impaired lung function, nonresponse to bronchodilators, and association with smoke exposure in utero. This group was also strongly correlated with late-onset thereby suggesting a distinct disease entity[16].

In the clinical community there is often an erroneous assumption that transient and EVW are relatively "benign", and persistent and MTW more troublesome. However, data from Manchester cohort have shown that almost 70% of all inhaled corticosteroids prescribed in the first year of life were prescribed to children in the transient wheeze group, although this class accounted for only one third of children who wheezed in the first year[12]. This would suggest that early-life transient wheezers may have more severe wheezing in infancy compared to those children who wheeze early, but go on to become persistent wheezers, and does not support the notion that transient and EVW are benign. These children may also be at risk of developing COPD in adulthood.

3.4 Phenotypes of severe asthma

The challenges in treatment of severe childhood wheezing/asthma due to high rates of non-responders suggests that children in the most severe category are either not taking their treatment, are undertreated, or do not respond to the currently available treatments. Although quantitatively small, this group of children consume a high proportion of resources[51] and therefore represent a significant economic burden.

In the US Severe Asthma Research Program (SARP), Fitzpatrick et al[52] used hierarchical clustering to identify four subgroups of severe asthmatic children differing in age of onset, lung function, FeNO and medication use. All subgroups were atopic, however, there was large variation in the magnitude of atopy. It is of note that the definition of asthma severity proposed by the ERS/ATS task force[53] was not fully applicable to any of the subgroups, despite initially using an ATS definition. In contrast, a study in Sweden did not find an association between atopy and severity of asthma[54]. They found that environmental factors, such as smoking, were significant risk factors for the severe asthma phenotype.

The TAP group identified two other severe phenotypes, though within a more heterogeneous asthmatic population: ‘asthma with severe exacerbations and multiple allergies’, and ‘severe asthma with bronchial obstruction’[55]. The first phenotype showed marked eosinophilic and basophilic inflammation, but the baseline FEV_1 values were within normal range, though slightly lower than those for the mild asthmatic group. This is consistent with the observation that children can exacerbate frequently despite high medication use and yet still have relatively normal FEV_1. The second phenotype was characterised by neutrophilia and high BMI. Additionally, this group of children had poorer lung function, though FEV_1 was still within normal limits as depicted by guidelines, confirming that spirometry in children may not be a good marker of severe childhood asthma.

3.5 Use of biomarkers to identify phenotypes

The heterogeneity of asthma with regards to symptom expression, response to therapy, and lung function has prompted the search for objective markers with better diagnostic value which are able to distinguish distinct subtypes of asthma. The gold standard for assessing the extent of airway inflammation is through the use of bronchial biopsies, bronchoalveolar lavage, and sputum induction. These techniques, however, are too invasive for children, and there has to be a clear clinical indication with intended benefit for this method of assessment to be employed[56]. As a result, biomarkers in childhood asthma/wheezing have mostly been limited to serum/blood and breath. These biomarkers primarily assess eosinophilic or Th2 inflammation[57].

3.5.1 Fractional concentration of exhaled Nitric Oxide (FeNO)

Fractional concentration of Nitric Oxide has been widely used in children as it is non-invasive and relatively simple. Several studies have shown strong correlations between FeNO levels and blood/sputum eosinophils[58], IgE[59-61], and serum eosinophilic cationic protein[62,63], with higher levels of FeNO denoting Th2-mediated inflammation and good response to inhaled corticosteroids[64]. 'Low' FeNO levels are thought to represent non-eosinophillic asthma subtypes that are less likely to respond to ICS treatment[65]. The optimal threshold to differentiate what constitutes 'high' or 'low' levels of FeNO is still up for debate. The most recent guidelines suggest FeNO <20ppb indicates poor response to inhaled corticosteroids, while FeNO >35ppb indicates good response[66], but that serial measurements should be taken over time for monitoring. However, this recommendation is based on studies that had different cut off points and different positive/negative predictive values with respect to inhaled steroid effect. It is also important to keep in mind that FeNO levels reflect inflammation and not necessarily asthma, and so interpretation of results should be taken within context, particularly in the case of atopic comorbidities. As it stands, rather than identifying subtypes of wheezing, the clinical utility of FeNO is currently limited to that of a screening method for eosinophilic and Th2 mediated inflammation with some indication of steroid response, and may be used as a marker of adherence with ICS treatment.

3.5.2 Exhaled Breath Condensate

Exhaled breath condensate (EBC), thought to mirror airway lining fluid[67], is obtained by cooling exhaled air from children breathing normally for 10-15 minutes[68]. Potential biomarkers include pH, markers of oxidative stress (8-isoprostane, hydrogen peroxide, aldehydes) and airway inflammation (eicosanoids, cytokines). Despite the lack of a standardised method, reference values, or proper validation[69], some interesting results have emerged. For example, 8-isoprostane has been found in increased quantities in both problematic and atopic asthmatic children despite ICS treatment[70-71]. Th2 Cytokines, such as IL-4, are also increased in asthmatic children, especially those with concurrent atopy[72] and persistent wheeze[73]. This cytokine was also found to be a good marker for assessing asthma control[72]. Additionally, cytokine IL-5 has been shown to predict asthma exacerbations[74]. However, other studies have questioned the utility of EBC (in particular measuring EBC pH) in childhood asthma[75].

3.5.3 Periostin

Periostin is an emerging biomarker in Th2 inflammation in studies in adults. It is induced by IL-4 and IL-13 in airway epithelial cells and lung fibroblasts and thought to mediate collagen synthesis and fibrillogenesis[76,77]. It has been suggested that high periostin is associated with good clinical response to corticosteroids in Th2 inflammation[78]. Recently, it has been used in clinical trials as a predictor of response to lebrikizumab (anti IL-13)[79]. However, periostin levels are higher in children compared to those in adults and change developmentally (likely due to bone turnover and growth), and it remains unclear whether periostin will be of use in phenotyping childhood wheezing and asthma.

3.6 Asthma phenotypes and genetic studies

Approaches for discovering genes connected to asthma have ranged from candidate gene association studies and genome-wide linkage studies, to more recent genome-wide association studies (GWAS). Several large studies (e.g. European GABRIEL consortium[80] and US EVE consortium[81]) identified a number of genetic associates of asthma in children (in particular in the region 17q21). However, the predictive value of these is low, with specificity being 75% and sensitivity only 35%. Other genes (IL33, IL1RL1, IL18R1) have also been implicated in asthma development[82]. Most genetic studies defined "asthma" as parentally or patient-reported "doctor-diagnosed asthma". Unless genetic studies find better ways to distinguish between different asthma subtypes under this umbrella diagnosis at a population level, it will be difficult (if not impossible) to discover their unique underlying genetic risk factors, as any signal will be diluted by phenotypic heterogeneity[83]. It is worth emphasizing that when a much more precise and specific definition of early-life onset asthma with recurrent, severe exacerbations in pre-school age was used in a GWAS, identified associations had a considerably greater effect size (e.g. in 17q21 region), and novel susceptibility genes such as *CDHR3* (cadherin-related family member 3, rs6967330, C_{529}Y]) were identified[84]. Subsequent studies have shown that CDHR3 may be a receptor for rhinovirus-C, and that the same genetic variant which was linked with hospitalizations for early-onset childhood asthma in birth cohort studies also mediates enhanced RV-C binding and replication[85]. These studies provide indirect evidence that we need to move away from using umbrella term of "doctor-diagnosed asthma" in genetic studies, and that investigation of genetic associates of different wheeze phenotypes may render more informative findings. As an example, the ALSPAC investigators extended genetic association studies to include wheeze phenotypes which they identified previously, and found that 17q21 locus was associated with persistent and intermediate wheeze, but not with transient wheeze, atopy or lung function[86]. Combining data from the ALSPAC and the PIAMA cohort, Savenije et al found that IL1RL1 and IL33 SNPs were associated with intermediate and late-onset wheeze and that this was in the presence of early sensitization, thus suggesting that allergic sensitization, through the IL33-IL1RL1 pathway, may be the key driver to wheeze and subsequent asthma development[87].

A different approach using transcriptomic analysis of peripheral blood in children with viral induced wheeze and exacerbations has shown that there are distinct microRNA profiles in CD8+ T cells, with reduced regulation of microRNA146a/b and microRNA28-5p[88].

Although we have been able to enhance our knowledge of genetic associations of wheeze phenotypes and asthma at a population level, we have still not been able to translate this knowledge to an individual level. This is partly due to the heterogeneous nature of childhood wheezing and asthma described above, but is also a consequence of gene-environment interactions whereby the genetic effect on asthma may be modified by certain environmental factors. Classical examples of the role of gene-environment interaction in modifying disease is finding that the impact of environmental endotoxin exposure on non-atopic wheeze and atopic sensitization was dependent on and modified by genetic variants in *CD14*[89,90], and studies demonstrating that the effect of day care on atopic wheezing is opposite among children with different variants of the *TLR2* gene[91]. The approach to understanding asthma within the complex context of gene-environment interactions may be an incentive utilise a candidate gene approach based on sound epidemiological principles of causality[92].

3.7 Conclusion

3.7.1 Lack of cohesive methodology for understanding asthma phenotypes in childhood

There is clear lack of standardisation in defining childhood wheezing illness and asthma which is leading to the inconsistent findings with respect to the genetic, environmental and physiological risk factors associated with this disease. In the last few years, a plethora of divergent statistical methods have been used in an attempt to understand the natural history of childhood wheezing illness. Although, by definition, all these techniques find patterns within a dataset and categorise them accordingly, there is no unified statistical method, leading to inconsistency in results of studies using different data-driven approaches. This is exacerbated by the problem of inconsistent use of variables in different analyses and inconsistencies in the data collection process itself. Indeed, this could explain the sizeable variation in the proportion of children assigned to these wheeze phenotypes among published studies (see Belgrave et al[83] for table breakdown). This may be a reflection of both the different time points at which data was collected and the inherent symptom patterns among the dataset due to differences in population. Therefore, results from statistical and machine learning approaches need to be interpreted within a larger clinical context, and it needs to be emphasized that although such an approach can be used to generate hypotheses, it remains an exploratory rather than confirmatory approach to understanding subtypes of disease, and therefore clinical interpretation and external epidemiological validation is essential.

The large quantity and ready availability of data, while posing many challenges, also presents an opportunity to understand underlying disease heterogeneity with greater certainty. This is because a quantitatively increased volume of data gives us the ability to distinguish between “data signal” from “noise due to random variation” more accurately. Advances in computational power facilitate data visualisation and pattern recognition in high-dimensional data with greater speed and accuracy. Combined with increased computational power and advances in statistical methods, the emerging field of systems biology, which combines biomarkers, genomics, metabolomics, proteomics and computational mathematics, may enable better identification of pathophysiology underpinning disease subtypes. Integrating all these components from different cohorts and datasets, while including developmental patterns of symptoms, may initially identify common characteristics on a wider population level, with the ultimate goal of better understanding the disease on an individual level.

3.7.2 Clinical implications of defining asthma phenotypes in childhood

There has been considerable progress in our understanding of the heterogeneity of asthma. Phenotyping asthma has started to provide a framework for further research into subtyping based on specific mechanisms. What has been shown here is that asthma can no longer be considered a single disease. Phenotyping in children can be either based on age of onset, triggers, severity, or symptoms. The clinical relevance of this would then be seen taking it one step further and analysing how each child’s disease pattern develops over time and tailoring treatment accordingly. This is in stark contrast to the current approach where it is assumed that one drug treats all.

Classifying children into groups of similar observable characteristics and features will enable better understanding and subsequent classification into subtypes that represent the genetic, environmental, and biomarker make-up of the disease. Once this has been achieved, we can then follow-up longitudinally over time in order to assess stability, remission, and new onset of disease.

The ultimate aim is to be one step closer to personalised medicine which would provide early prevention and/or tailored treatment leading to more efficient use of resources and more efficacious control of disease in children. However, we are still not ready to apply this to every day clinical practice.

3.8 References

1. Bjerg, A., et al., *Time trends in asthma and wheeze in Swedish children 1996-2006: prevalence and risk factors by sex.* Allergy, 2010. **65**(1): p. 48-55.
2. Ronmark, E., et al., *The Obstructive Lung Disease in Northern Sweden (OLIN) longitudinal paediatric study I--the first 10 years.* Clin Respir J, 2008. **2 Suppl 1**: p. 26-33.
3. Belgrave, D., A. Simpson, and A. Custovic, *Challenges in interpreting wheeze phenotypes: the clinical implications of statistical learning techniques.* Am J Respir Crit Care Med, 2014. **189**(2): p. 121-3.
4. Wenzel, S.E., *Asthma phenotypes: the evolution from clinical to molecular approaches.* Nat Med, 2012. **18**(5): p. 716-25.
5. Anderson, G.P., *Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease.* Lancet, 2008. **372**(9643): p. 1107-1119.
6. Lotvall, J., et al., *Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome.* J Allergy Clin Immunol, 2011. **127**(2): p. 355-60.
7. Custovic, A., et al., *The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together.* Thorax, 2015. **70**(8): p. 799-801.
8. Fleming, L., et al., *Sputum inflammatory phenotypes are not stable in children with asthma.* Thorax, 2012. **67**(8): p. 675-81.
9. Deliu, M., et al., *Identification of Asthma Subtypes Using Clustering Methodologies.* Pulm Ther, 2016. **2**: p. 19-41.
10. Howard, R., et al., *Distinguishing Asthma Phenotypes Using Machine Learning Approaches.* Curr Allergy Asthma Rep, 2015. **15**(7): p. 38.
11. Martinez, F.D., et al., *Asthma and wheezing in the first six years of life. The Group Health Medical Associates.* N Engl J Med, 1995. **332**(3): p. 133-8.
12. Belgrave, D.C., et al., *Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing.* J Allergy Clin Immunol, 2013. **132**(3): p. 575-583 e12.
13. Henderson, J., 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): p. 974-80.
14. Collins, S.A., 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): p. 683-92.
15. Savenije, O.E., et al., *Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA.* J Allergy Clin Immunol, 2011. **127**(6): p. 1505-12 e14.
16. Depner, M., et al., *Clinical and epidemiologic phenotypes of childhood asthma.* American journal of respiratory and critical care medicine, 2014. **189**(2): p. 129-138.
17. Chen, Q., et al., *Using latent class growth analysis to identify childhood wheeze phenotypes in an urban birth cohort.* Annals of Allergy, Asthma & Immunology, 2012. **108**(5): p. 311-315.
18. Lowe, L.A., et al., *Wheeze phenotypes and lung function in preschool children.* Am J Respir Crit Care Med, 2005. **171**(3): p. 231-7.
19. Belgrave, D.C.M., et al., *Trajectories of lung function during childhood.* American journal of respiratory and critical care medicine, 2014. **189**(9): p. 1101-9.

20. Sears, M.R., et al., *A longitudinal, population-based, cohort study of childhood asthma followed to adulthood.* N Engl J Med, 2003. **349**(15): p. 1414-22.
21. Morgan, W.J., et al., *Outcome of asthma and wheezing in the first 6 years of life: follow-up through adolescence.* Am J Respir Crit Care Med, 2005. **172**(10): p. 1253-8.
22. Turner, S.W., et al., *Infants with flow limitation at 4 weeks: outcome at 6 and 11 years.* Am J Respir Crit Care Med, 2002. **165**(9): p. 1294-8.
23. Lau, S., et al., *Transient early wheeze is not associated with impaired lung function in 7-yr-old children.* Eur Respir J, 2003. **21**(5): p. 834-41.
24. Granell, R., J.A. Sterne, and J. Henderson, *Associations of different phenotypes of wheezing illness in early childhood with environmental variables implicated in the aetiology of asthma.* PLoS One, 2012. **7**(10): p. e48359.
25. Stein, R.T. and F.D. Martinez, *Asthma phenotypes in childhood: lessons from an epidemiological approach.* Paediatr Respir Rev, 2004. **5**(2): p. 155-61.
26. Scott, M., et al., *Understanding the nature and outcome of childhood wheezing.* Eur Respir J, 2009. **33**(3): p. 700-1.
27. Kurukulaaratchy, R.J., et al., *Are influences during pregnancy associated with wheezing phenotypes during the first decade of life?* Acta Paediatr, 2005. **94**(5): p. 553-8.
28. Henderson, J., 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): p. 974-80.
29. Stern DA, M.W., Taussig LM, Wright AL, Halonen M, Martinez FD, *Lung function at age 11 in relation to early wheezing.* American journal of respiratory and critical care medicine, 1999. **195**: p. A148.
30. Phelan, P.D., C.F. Robertson, and A. Olinsky, *The Melbourne Asthma Study: 1964-1999.* J Allergy Clin Immunol, 2002. **109**(2): p. 189-94.
31. Belgrave, D.C.M., et al., *Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing.* The Journal of allergy and clinical immunology, 2013. **132**(3): p. 575-583.e12.
32. Andersson, M., et al., *Remission and persistence of asthma followed from 7 to 19 years of age.* Pediatrics, 2013. **132**(2): p. e435-42.
33. Simpson, A., et al., *Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study.* American journal of respiratory and critical care medicine, 2010. **181**(11): p. 1200-1206.
34. Lazic, N., et al., *Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts.* Allergy, 2013. **68**(6): p. 764-70.
35. Garden, F.L., et al., *Atopy phenotypes in the Childhood Asthma Prevention Study (CAPS) cohort and the relationship with allergic disease: clinical mechanisms in allergic disease.* Clin Exp Allergy, 2013. **43**(6): p. 633-41.
36. Sherriff, A., et al., *Risk factor associations with wheezing patterns in children followed longitudinally from birth to 3(1/2) years.* Int J Epidemiol, 2001. **30**(6): p. 1473-84.
37. Brand, P.L., et al., *Definition, assessment and treatment of wheezing disorders in preschool children: an evidence-based approach.* Eur Respir J, 2008. **32**(4): p. 1096-110.
38. Just, J., et al., *Wheeze phenotypes in young children have different courses during the preschool period.* Ann Allergy Asthma Immunol, 2013. **111**(4): p. 256-261 e1.

39. Spycher, B.D., et al., *Distinguishing phenotypes of childhood wheeze and cough using latent class analysis.* Eur Respir J, 2008. **31**(5): p. 974-81.

40. Schultz, A. and P.L. Brand, *Episodic viral wheeze and multiple trigger wheeze in preschool children: a useful distinction for clinicians?* Paediatr Respir Rev, 2011. **12**(3): p. 160-4.

41. Caudri, D., et al., *Predicting the long-term prognosis of children with symptoms suggestive of asthma at preschool age.* J Allergy Clin Immunol, 2009. **124**(5): p. 903-10 e1-7.

42. Tromp, II, et al., *Dietary patterns and respiratory symptoms in pre-school children: the Generation R Study.* Eur Respir J, 2012. **40**(3): p. 681-9.

43. Schultz, A., et al., *The transient value of classifying preschool wheeze into episodic viral wheeze and multiple trigger wheeze.* Acta Paediatr, 2010. **99**(1): p. 56-60.

44. Schultz, A. and P.L. Brand, *Phenotype-directed treatment of pre-school-aged children with recurrent wheeze.* J Paediatr Child Health, 2012. **48**(2): p. E73-8.

45. Just, J., et al., *Novel severe wheezy young children phenotypes: boys atopic multiple-trigger and girls nonatopic uncontrolled wheeze.* J Allergy Clin Immunol, 2012. **130**(1): p. 103-10 e8.

46. Kappelle, L. and P.L. Brand, *Severe episodic viral wheeze in preschool children: High risk of asthma at age 5-10 years.* Eur J Pediatr, 2012. **171**(6): p. 947-54.

47. Goksor, E., et al., *Asthma symptoms in early childhood--what happens then?* Acta Paediatr, 2006. **95**(4): p. 471-8.

48. Frank, P.I., et al., *Long term prognosis in preschool children with wheeze: longitudinal postal questionnaire study 1993-2004.* BMJ, 2008. **336**(7658): p. 1423-6.

49. Devulapalli, C.S., et al., *Severity of obstructive airways disease by age 2 years predicts asthma at 10 years of age.* Thorax, 2008. **63**(1): p. 8-13.

50. Matricardi, P.M., et al., *Wheezing in childhood: incidence, longitudinal patterns and factors predicting persistence.* Eur Respir J, 2008. **32**(3): p. 585-92.

51. Godard, P., et al., *Costs of asthma are correlated with severity: a 1-yr prospective study.* Eur Respir J, 2002. **19**(1): p. 61-7.

52. Fitzpatrick, A.M., et al., *Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program.* J Allergy Clin Immunol, 2011. **127**(2): p. 382-389 e1-13.

53. Chung, K.F., et al., *International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma.* Eur Respir J, 2014. **43**(2): p. 343-73.

54. Konradsen, J.R., et al., *Problematic severe asthma: a proposed approach to identifying children who are severely resistant to therapy.* Pediatr Allergy Immunol, 2011. **22**(1 Pt 1): p. 9-18.

55. Just, J., et al., *Two novel, severe asthma phenotypes identified during childhood using a clustering approach.* Eur Respir J, 2012. **40**(1): p. 55-60.

56. Vijverberg, S.J., et al., *Clinical utility of asthma biomarkers: from bench to bedside.* Biologics, 2013. **7**: p. 199-210.

57. Loutsios, C., et al., *Biomarkers of eosinophilic inflammation in asthma.* Expert Rev Respir Med, 2014. **8**(2): p. 143-50.

58. Malinovschi, A., et al., *Simultaneously increased fraction of exhaled nitric oxide levels and blood eosinophil counts relate to increased asthma morbidity.* J Allergy Clin Immunol, 2016.

59. Brussee, J.E., et al., *Exhaled nitric oxide in 4-year-old children: relationship with asthma and atopy.* Eur Respir J, 2005. **25**(3): p. 455-61.
60. Romero, K.M., et al., *Role of exhaled nitric oxide as a predictor of atopy.* Respir Res, 2013. **14**: p. 48.
61. Scott, M., et al., *Influence of atopy and asthma on exhaled nitric oxide in an unselected birth cohort study.* Thorax, 2010. **65**(3): p. 258-62.
62. Mogensen, I., et al., *Simultaneously elevated exhaled nitric oxide and serum-ECP relate to recent asthma events in asthmatics in a cross sectional population based study.* Clin Exp Allergy, 2016.
63. Warke, T.J., et al., *Exhaled nitric oxide correlates with airway eosinophils in childhood asthma.* Thorax, 2002. **57**(5): p. 383-7.
64. Mahr, T.A., J. Malka, and J.D. Spahn, *Inflammometry in pediatric asthma: a review of fractional exhaled nitric oxide in clinical practice.* Allergy Asthma Proc, 2013. **34**(3): p. 210-9.
65. Ludviksdottir, D., et al., *Clinical aspects of using exhaled NO in asthma diagnosis and management.* Clin Respir J, 2012. **6**(4): p. 193-207.
66. Dweik, R.A., et al., *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): p. 602-15.
67. Hunt, J., *Exhaled breath condensate: an evolving tool for noninvasive evaluation of lung disease.* J Allergy Clin Immunol, 2002. **110**(1): p. 28-34.
68. Baraldi, E. and S. Carraro, *Exhaled NO and breath condensate.* Paediatr Respir Rev, 2006. **7 Suppl 1**: p. S20-2.
69. Loukides, S., et al., *Exhaled breath condensate in asthma: from bench to bedside.* Curr Med Chem, 2011. **18**(10): p. 1432-43.
70. Carraro, S., et al., *EIA and GC/MS analysis of 8-isoprostane in EBC of children with problematic asthma.* Eur Respir J, 2010. **35**(6): p. 1364-9.
71. Baraldi, E., et al., *Cysteinyl leukotrienes and 8-isoprostane in exhaled breath condensate of children with asthma exacerbations.* Thorax, 2003. **58**(6): p. 505-9.
72. Robroeks, C.M., et al., *Exhaled nitric oxide and biomarkers in exhaled breath condensate indicate the presence, severity and control of childhood asthma.* Clin Exp Allergy, 2007. **37**(9): p. 1303-11.
73. van de Kant, K.D., et al., *Elevated inflammatory markers at preschool age precede persistent wheezing at school age.* Pediatr Allergy Immunol, 2012. **23**(3): p. 259-64.
74. Robroeks, C.M., et al., *Prediction of asthma exacerbations in children: results of a one-year prospective study.* Clin Exp Allergy, 2012. **42**(5): p. 792-8.
75. Nicolaou, N.C., et al., *Exhaled breath condensate pH and childhood asthma: unselected birth cohort study.* Am J Respir Crit Care Med, 2006. **174**(3): p. 254-9.
76. Sidhu, S.S., et al., *Roles of epithelial cell-derived periostin in TGF-beta activation, collagen production, and collagen gel elasticity in asthma.* Proc Natl Acad Sci U S A, 2010. **107**(32): p. 14170-5.
77. Norris, R.A., et al., *Periostin regulates collagen fibrillogenesis and the biomechanical properties of connective tissues.* J Cell Biochem, 2007. **101**(3): p. 695-711.
78. Woodruff, P.G., et al., *Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids.* Proc Natl Acad Sci U S A, 2007. **104**(40): p. 15858-63.

79. Corren, J., et al., *Lebrikizumab treatment in adults with asthma.* N Engl J Med, 2011. **365**(12): p. 1088-98.

80. Moffatt, M.F., et al., *A large-scale, consortium-based genomewide association study of asthma.* N Engl J Med, 2010. **363**(13): p. 1211-21.

81. Torgerson, D.G., et al., *Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations.* Nat Genet, 2011. **43**(9): p. 887-92.

82. Grotenboer, N.S., et al., *Decoding asthma: translating genetic variation in IL33 and IL1RL1 into disease pathophysiology.* J Allergy Clin Immunol, 2013. **131**(3): p. 856-65.

83. Belgrave, D.C., A. Custovic, and A. Simpson, *Characterizing wheeze phenotypes to identify endotypes of childhood asthma, and the implications for future management.* Expert Rev Clin Immunol, 2013. **9**(10): p. 921-36.

84. Bonnelykke, K., et al., *A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations.* Nat Genet, 2014. **46**(1): p. 51-5.

85. Bochkov, Y.A., et al., *Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication.* Proc Natl Acad Sci U S A, 2015. **112**(17): p. 5485-90.

86. Granell, R., et al., *Examination of the relationship between variation at 17q21 and childhood wheeze phenotypes.* J Allergy Clin Immunol, 2013. **131**(3): p. 685-94.

87. Savenije, O.E., et al., *Association of IL33-IL-1 receptor-like 1 (IL1RL1) pathway polymorphisms with wheezing phenotypes and asthma in childhood.* J Allergy Clin Immunol, 2014. **134**(1): p. 170-7.

88. Tsitsiou, E., et al., *Transcriptome analysis shows activation of circulating CD8+ T cells in patients with severe asthma.* J Allergy Clin Immunol, 2012. **129**(1): p. 95-103.

89. Simpson, A., et al., *Endotoxin exposure, CD14, and allergic disease: an interaction between genes and the environment.* Am J Respir Crit Care Med, 2006. **174**(4): p. 386-92.

90. Simpson, A. and F.D. Martinez, *The role of lipopolysaccharide in the development of atopy in humans.* Clin Exp Allergy, 2010. **40**(2): p. 209-23.

91. Custovic, A., et al., *Effect of day care attendance on sensitization and atopic wheezing differs by Toll-like receptor 2 genotype in 2 population-based birth cohort studies.* J Allergy Clin Immunol, 2011. **127**(2): p. 390-397 e1-9.

92. Tabor, H.K., N.J. Risch, and R.M. Myers, *Candidate-gene approaches for studying complex genetic traits: practical considerations.* Nat Rev Genet, 2002. **3**(5): p. 391-7.

Part 2: Clinical Application

Chapter 4 Features of asthma which provide meaningful insights for understanding the disease heterogeneity

Matea Deliu, Tolga S. Yavuz, Matthew Sperrin, Danielle Belgrave, Umit Sahiner, Cansin Sackesen, Omer Kalayci, Adnan Custovic

Clin Exp Allergy. 2017 Aug 22. doi: 10.1111/cea.13014

4.1 Rationale for the study

The rationale for this study was to analyse a data-rich cross-sectional cohort in order to ascertain clinical features of asthma subtypes. The dataset has very little missing data and so removes the need to impute, thereby providing robust and stable results. Although cross-sectional studies are limited in their ability to draw conclusions about causality or association since the risk factors and outcomes are measured at the same time point, they are good starting points for descriptive analytics and generating hypotheses. With that, this chapter serves as a framework for ascertaining key features of disease that are then expanded upon in the subsequent chapters.

4.2 Abstract

Background: Data-driven methods such as hierarchical clustering (HC) and principal component analysis (PCA) have been used to identify asthma subtypes, with inconsistent results.

Objective: To develop a framework for the discovery of stable and clinically meaningful asthma subtypes.

Methods: We performed HC in a rich dataset from 613 asthmatic children, using 45 clinical variables (Model 1), and after PCA dimensionality reduction (Model 2). Clinical experts then identified a set of asthma features/domains which informed clusters in the two analyses. In Model 3, we re-clustered the data using these features to ascertain whether this improved the discovery process.

Results: Cluster stability was poor in Models 1 and 2. Clinical experts highlighted four asthma features/domains which differentiated the clusters in two models: age of onset, allergic sensitization, severity, and recent exacerbations. In Model 3 (HC using these four features), cluster stability improved substantially. The cluster assignment changed, providing more clinically interpretable results. In a 5-cluster model, we labelled the clusters as: "Difficult asthma" (n=132); "Early-onset mild atopic" (n=210); "Early-onset mild non-atopic: (n=153); "Late-onset" (n=105); and "Exacerbation-prone asthma" (n=13). Multinomial regression demonstrated that lung function was significantly diminished among children with "Difficult asthma"; blood eosinophilia was a significant feature of "Difficult", "Early-onset mild atopic", and "Late-onset asthma". Children with moderate-severe asthma were present in each cluster.

Conclusions and clinical relevance: An integrative approach of blending the data with clinical expert domain knowledge identified four features, which may be informative for ascertaining asthma endotypes. These findings suggest that variables which are key determinants of asthma presence, severity or control, may not be the most informative for determining asthma subtypes. Our results indicate that exacerbation-prone asthma may be a separate asthma endotype, and that severe asthma is not a single entity, but an extreme end of the spectrum of several different asthma endotypes.

4.3 Introduction

The evidence is mounting that asthma is an umbrella diagnosis for a collection of distinct diseases (endotypes), with varying phenotypic expression of characteristic symptoms (ranging from wheezing and shortness of breath, to cough and chest tightness), and accompanying variable airflow obstruction.[1-3] It is important to make a clear distinction between asthma phenotypes (which are observable and measured characteristics of the disease)[9] and asthma endotypes (which is a term that refers to the subtype of the disease with a clearly defined underlying mechanism).[1,2,10] It is of note that similar symptoms and observable features can arise through different pathophysiological mechanisms, and that consequently different endotypes may have similar, or even the same phenotype.

Identifying true endotypes of asthma and their underlying mechanisms is a pre-requisite for achieving better mechanism-based treatment targeting, and ultimately delivery of genuinely stratified medicine in asthma.[10] However, although the current consensus in the medical community is that different asthma endotypes do exist, there is little agreement on what these are and how best to define them.[5]

Approaches utilized in the search for asthma endotypes have ranged from investigator-led pattern identification in the clinical setting, to supervised and unsupervised statistical modelling techniques that utilize large amounts of data and computer algorithms to find the latent (hidden, unknown *a-priori*) patterns of observable features (such as symptoms, medication use, allergic sensitization, lung function), either in cross-sectional studies[11-14] or over time. Data-driven approaches allow interrogation of data without imposing *a-priori* hypotheses, hence eliminating investigator bias and enabling novel hypotheses to be generated.[5] In most previous studies which used such approaches, the selection of variables used for subtype discovery was either pre-determined by clinical advice,[11,13,20] or by the use of statistical data reduction techniques such as principal component analysis (PCA).[12,21,22] Although valuable information has been gained, and there was some (but not complete) resemblance between the results, most studies reported different disease clusters; several recent reviews have summarized these findings.[6,16-18,23] These inconsistencies may be explained by the inherent heterogeneity among different populations, the differences in clustering techniques used, the lack of consistency in selecting variables, their encodings and transformations, or the use of excessive numbers of variables which may result in subtype 'signals' being drowned in the noise.[24]

When selecting the variables for unsupervised analyses, the investigators rely on the data which is available (e.g. in birth cohorts[14,15,19] or studies of adults and children with established disease[11-13]). In most clinical studies, the assessment and monitoring of study participants focuses on measures which aim to ascertain asthma presence, severity, control, and responsiveness to treatment. We hypothesise that these may not necessarily be the variables or features which are most informative for the discovery of disease endotypes. We propose that a careful synergy of data-driven methods and clinical interpretation may help us to better understand the heterogeneity of asthma and enable the discovery of true asthma endotypes. In this study, we aimed to ascertain whether a framework for data interrogation which utilises an integrative approach that brings together the data and bio-statistical expertise, with a clinical expert domain knowledge and clinical experience, can facilitate the identification of stable and clinically meaningful asthma subtypes.

4.4 Methods

4.4.1 Study design, setting and participants

We used anonymized data from a cross-sectional study which recruited children with asthma aged 6-18 years from two hospitals in Ankara, Turkey (Hacettepe and GATA University Hospitals); the study is described in detail elsewhere.[24-26] Briefly, children who presented to the Paediatric Allergy and Asthma Units completed skin prick tests, spirometry, and measurement of bronchodilator reversibility (BDR). Amongst children with a negative BDR test (<12% increase in FEV_1 following administration of 200 μg of albuterol), airway hyper-responsiveness (AHR) was assessed using methacholine or exercise challenge test.

Asthma was defined as all three of the following: (1) physician-diagnosed asthma; (2) current use of asthma medication; and (3) either BDR or AHR (positive methacholine or exercise challenge test). Children with other known systemic disorders such as cystic fibrosis or immunodeficiency, and those who had a severe exacerbation requiring systemic corticosteroids or hospital admission within the previous four weeks were not included.

4.4.2 Data sources/measurements

We recorded a total of 47 variables for each study participant; of those, 45 were used in the analysis (Table E4.1).

Symptoms, exacerbations and prescribed medications: A modified ISAAC questionnaire was interviewer-administered to ascertain the age of onset, the presence of asthma-related symptoms within the past 4 weeks, the number of asthma exacerbations within the past year, and hospitalizations for acute asthma (ever).

Asthma severity: Categorised as mild, moderate or severe based on GINA guidelines (www.ginasthma.org); a detailed description is published elsewhere.[25] Briefly, we allocated patients to severity group based on the assessment of clinical symptoms before the treatment was initiated; when the patient was already receiving treatment, the severity was assigned based on the clinical features and the step of the daily medication regimen (for details, please see Online supplement).

Lung function: We performed spirometry, methacholine and/or exercise challenge tests according to ATS/ERS guidelines;[29,30] FEV_1 (% predicted), FVC, FEV_1/FVC and FEF_{25-75}, were recorded.[32,33]

Allergic sensitization: We carried out skin prick testing to a battery of allergens including dust mite, tree, grass and weed pollens, moulds, cat, dog, cockroach and horse. Wheal 3 mm greater than negative control was considered a positive reaction. We also measured total serum IgE.

Objective measurements: Height, weight, body mass index (BMI; standardized for age and growth and sex), and blood eosinophils.

4.4.3 Statistical methods

All analyses were performed in R software (www.r-project.org/).[34] For a detailed description of statistical methods please see the online supplement. Briefly, we performed a hierarchical cluster analysis (HC) using three different models:

1. HC after PCA dimensionality reduction: We first performed PCA on all variables in the dataset, and then carried out HC using principal components with eigenvalues>1.

2. HC using all available variables: We performed HC on raw data, without removing or modifying any of the variables.

3. Identification of a subset of potentially important features, and clustering using the informative subset: The results of the first two models were reviewed by clinical experts to identify features (domains) in the data set which may drive cluster allocation. We then used these informative features in a further HC.

Cluster stability was tested with bootstrapping methods. The data were resampled and the Jaccard similarities of the original clusters to the most similar clusters in the resampled data were computed. The mean of the similarities were used as an index of stability, and a mean greater than 0.75 was deemed as stable.[35]

We used logistic regression to identify variables which differed between the clusters.

All study procedures were done in accordance with a protocol previously approved by the Ethics Committee of Hacettepe University Ethics committee (# FON 02/24-1) and the Ethics Committee of Gulhane School of Medicine (05.06.2013/21). All parents provided written informed consent and children provided assent for the study procedures.

4.5 Results

4.5.1 Participants and descriptive data

The study population comprised of 613 asthmatic children (64% male, median age 9 years, 49% with physician-diagnosed allergic rhinitis, 39% exposed to tobacco smoke, 59% atopic, all receiving SABA as needed, 61% receiving ICS, 15% experiencing 2 or more asthma exacerbations in the previous year, with mean FEV_1 % predicted of 87%). The characteristics of the study population are shown in Table 4.1. Asthma was classified as mild, moderate, or severe in 78%, 20%, and 2% of cases, respectively.

Table 4.1: *Demographic characteristics of the study population. Definition of abbreviations: BMI = body mass index, FEV1 = forced expiratory volume in 1 second, FVC = forced vital capacity, *ICS = inhaled corticosteroid dose represented as BDP equivalent, LABA = long acting beta$_2$-agonist, SABA = short acting beta$_2$-agonist. Continuous variables are given as mean and standard deviation, binary variables are given as percentages with absolute values*

N=613	**Mean (SD)** **% (N)**
Age at follow-up (years)	9 (3.0)
Sex (male)	64% (392)
BMI	18.4 (3.6)
Age of asthma onset (years)	5 (3.4)
Family history of asthma (yes)	30% (184)
Exposure to tobacco smoke (yes)	39% (240)
Skin prick test positivity	59% (361)
FEV_1 % predicted	87 (14.3)

FVC % predicted	96 (15.1)
FEV$_1$/FVC (%)	86 (7.0)
Bronchodilator Reversibility (%)	17.1 (12.9)
Total IgE (kU/L)	228 (458)
Blood eosinophil (%)	4.4 (3.5)
Asthma Severity	
Mild	78% (476)
Moderate	20% (126)
Severe	2% (11)
Using regular ICS	61% (375)
ICS dose >400mcg*	18% (113)
Using regular Montelukast	8% (51)
Using regular controller medication (ICS/LABA and/or Montelukast)	63% (385)
Using regular ICS/LABA	8% (51)
2 or more asthma attacks within the last year	15% (95)
2 or more hospitalizations for asthma ever	5% (29)
Presence of rhinitis	49% (302)
Presence of eczema	6% (37)

4.5.2 Data-driven analyses: Dimensionality reduction vs. clustering using all available variables

HC after dimensionality reduction: Dimensionality reduction using PCA identified 19 components with eigenvalues above 1, which accounted for 73% of the variance within the dataset. The correlation matrix of the variables is shown in Figure E4.1. Variables describing atopy correlated highly, as did those relating to lung function and medication use. Table

E4.2 shows the eigenvalues and variance explained by 19 components, and Table E4.3 the variable contribution/loading to each of the first five components.

A five-cluster model in HC after PCA dimensionality reduction provided the most clinically interpretable results. Table E4.4 shows clinical features/variables which differed across the clusters. Based on their dominant features, we labelled the clusters as: Cluster 1 (n=102), "Moderately-severe asthma with poor lung function, high symptom burden and medication use"; Cluster 2 (n=70), "Middle school-age onset, predominantly male, with high symptom burden despite normal lung function"; Cluster 3 (n=117), "Late-onset, multiple sensitization, mild asthma with diminished lung function"; Cluster 4 (n=149), "Early-onset atopic mild asthma, predominantly female"; Cluster 5 (n=175), "Mild atopic asthma". Children in Cluster 1 had the lowest lung function, with FEV_1 21% lower compared to those in Cluster 5. Clusters 2 and 3 comprised of predominantly boys, while those in Cluster 4 were mostly girls. Allergic comorbidities were significant features of Cluster 3.

HC using all available variables: As in the previous model, a five-cluster solution provided the most clinically interpretable results. However, the clusters were different, both in terms of clinical characteristics and the number of children in each cluster. Table E5 highlights clinical features and variables which differed across the clusters. We labelled the clusters as: Cluster 1 (n=168), "Early-onset severe asthma, predominantly female"; Cluster 2 (n=100), "Late-onset mild atopic asthma"; Cluster 3 (n=103), "Moderate-severe atopic asthma"; Cluster 4 (n=223), "Mild non-atopic asthma, predominantly male"; Cluster 5 (n=19), "Middle-school age of onset, atopic, with frequent exacerbations". Children in Cluster 3 had the poorest lung function (mean FEV_1 72.6%), Cluster 2 was associated with allergic comorbidities, and Cluster 5 was predominantly associated with exacerbations (Table E4.5).

Cluster stability: Cluster stability was generally poor for both models, with HC on principal components producing only one stable cluster (Cluster 1), and HC using all available data producing two stable clusters (Clusters 2 and 5).

4.5.3 Blending the data and bio-statistical expertise with clinical expert domain knowledge

Identification of stable features which distinguish the clusters: We first compared the subject allocation between the two analyses to ascertain the overlap which could indicate similarity (Table E4.6). However, there was little overlap (apart from one cluster pair, Cluster 5 in HC after PCA, and Cluster 4 in HC using all variables). We therefore proceeded with the comparison of the characteristics of clusters which we identified using the two methods. Clinical domain experts reviewed the results (Tables E4.4-E4.6) to highlight features and variables which characterized each cluster, and similarities and differences between the clusters (Table E4.7). We then used clinical expert domain knowledge and experience to identify four disease features/domains common to each cluster in both models: 1. Age of onset; 2. Allergic sensitization; 3. Asthma severity; and 4. Recent exacerbations. We assigned these four features as an "informative set", and proceeded to ascertain whether using this set may help distinguish asthma subtypes.

HC using the informative set of features: In HC using this informative subset of features, a five-cluster solution provided the most clinically interpretable results. Compared to previous analyses, the cluster assignment changed, but the cluster stability improved substantially (Table E4.8, bootstrap mean$\geq$0.99). Table 4.2 shows clinical features which differed across the clusters. Based on the dominant features of each cluster, we labelled them as: Cluster 1 (n=132), "Difficult asthma"; Cluster 2 (n=210), "Early-onset mild atopic asthma"; Cluster 3 (n=153), "Early-onset mild non-atopic asthma"; Cluster 4 (n=105), "Late-onset asthma"; and Cluster 5 (n=13), "Exacerbation-prone asthma".

By varying the definition of allergic sensitization from the dichotomous (sensitized/not sensitized; Table 4.2), to ordinal (non-atopic, mono-sensitized, poly-sensitized; Table E4.9) and continuous (IgE titre; Table E4.10), we found that the clusters remained very similar despite some changes to cluster allocation. However, the cluster stability slightly decreased.

We validated the clusters in relation to lung function (FEV$_1$, FEV$_1$/FVC, BDR), blood eosinophils, allergic comorbidities (eczema or rhinitis), family history and environmental exposures (Table 4.3). Multinomial regression model using children in Cluster 3 (with mildest asthma) as the reference has indicated that lung function was significantly diminished only among children in Cluster 1 ("Difficult asthma"). High blood eosinophilia was a significant feature of "Difficult asthma", "Early-onset mild atopic asthma" and "Late-onset asthma" clusters, while family history of asthma and concurrent rhinitis were most common among children in "Early-onset mild atopic asthma" cluster. Exposure to tobacco smoke was highest among children in the “Difficult asthma” cluster, although this did not reach statistical significance (p=0.09). There was no difference in pet ownership and eczema between the clusters. Children with moderate/severe asthma were present in each of the clusters (Cluster 1, 65%; Cluster 2, 10%; Cluster 3, 8%; Cluster 4, 13%; Cluster 5, 38%).

Table 4.2: *Univariate logistic regression analysis showing the clinical features that differed across the clusters derived by HC using the four informative features/domains (dichotomous definition of allergic sensitization).Quantitative variables are represented as mean (95% CI). Ordinal variables are represented as proportions (%).*Coeff: The coefficient translates into a value of how likely a child is assigned to that cluster based on the variable response.*

Feature/domain	**Cluster 1 (n=132)** *"Difficult asthma"*		**Cluster 2 (n=210)** *"Early-onset mild atopic asthma"*		**Cluster 3 (n=153)** *"Early-onset mild non-atopic asthma"*		**Cluster 4 (n=105)** *"Late-onset asthma"*		**Cluster 5 (n=13)** *"Exacerbation-prone asthma"*	
	Coeff* Mean (95%CI) or frequency (%)	P-value	Coeff Mean (95%CI) or frequency (%)	P-value	Coeff Mean (95%CI) or frequency (%)	P-value	Coeff Mean (95%CI) or frequency (%)	P-value	Coeff Mean (95%CI) or frequency (%)	P-value
Age of Onset	**-0.03**	**0.04**	**-0.10**	**<0.001**	**-0.12**	**<0.001**	**0.26**	**<0.001**	-0.008	0.14
Years	4.9 (2.3-7)		4.4 (3-6)		3.8 (2-6)		10.7 (9-12)		4.1 (2-5)	
Asthma attacks	0.008	0.96	**-0.04**	**0.04**	**-0.04**	**0.02**	**-0.04**	**0.007**	**0.12**	**<0.001**
Number, previous year	1.0 (0-1)		**0.8 (0-1)**		**0.9 (0-1)**		**0.4 (0-1)**		**3.5 (0-7)**	
Allergic sensitization	-0.002	0.88	**0.29**	**<0.001**	**-0.29**	**<0.001**	0.01	0.26	-0.002	0.71
Sensitized	77/132 (58%)		183/210 (87%)		27/153 (18%)		67/105 (64%)		7/13 (54%)	
Asthma Severity	**0.38**	**<0.001**	**-0.17**	**<0.001**	**-0.13**	**<0.001**	**-0.08**	**<0.001**	0.006	0.27
Mild	46/132 (35%)		190/210 (90%)		141/153 (92%)		91/105 (87%)		8/13 (62%)	
Moderate/severe	86/132(65%)		20/210 (10%)		12/153 (8%)		14/105 (13%)		5/13 (38%)	

Cluster stability	**1.00**	**0.99**	**0.99**	**1.00**	**1.00**

Table 4.3*: Multinomial logistic regression analysis showing lung function, blood eosinophils, tobacco smoke exposure, pet ownership, family history of asthma, and comorbidities across the five clusters derived by HC using the four informative features/domains. Quantitative variables are represented as mean (95% CI). Ordinal variables are represented as proportions (%); RR: relative risk, CI: confidence interval*

		Cluster 3 (n=153) *"Early-onset mild non-atopic asthma"*	**Cluster 1 (n=132)** *"Difficult asthma"*		**Cluster 2 (n=210)** *"Early-onset mild atopic asthma"*		**Cluster 4 (n=105)** *"Late-onset asthma"*		**Cluster 5 (n=13)** *"Exacerbation-prone asthma"*	
FEV_1 % predicted	Mean (95%CI)	88.4 (86-91)	83.0 (74-91)		88.0 (80-96)		87.9 (78-97)		83.2 (74-90)	
	RR (95% CI)	N/A (Reference group)	**0.68 (0.53-0.86)**	**P<0.001**	0.97 (0.79-1.20)	P=0.82	0.85 (0.75-1.25)	P=0.81	0.82 (0.39-1.22)	P=0.19
FEV_1/FVC (%)	Mean (95%CI)	86.6 (85.5-87.7)	84.8 (83.5-86.1)		86.3 (85.4-87.2)		85.4 (84.1-86.8)		83.7 (78.8-88.5)	
	RR (95% CI)	N/A (Reference group)	**0.77 (0.61-1.09)**	**P=0.03**	0.95 (0.77-1.18)	P=0.65	0.83 (0.65-1.08)	P=0.17	0.67 (0.39-1.14)	P=0.14
Bronchodilator reversibility (BDR), %	Mean (95%CI)	16.5 (14.7-18.4)	18.9 (17.6-20.2)		16.6 (14.8-18.4)		17.5 (14.6-20.5)		12.5 (9.4-15.6)	
	RR (95% CI)	N/A (Reference group)	1.18 (0.93-1.49)	P=0.16	1.00 (0.79-1.36)	P=0.98	1.08 (0.83-1.39)	P=0.55	0.58 (0.25-1.37)	P=0.22
Blood eosinophils, %	Mean (95%CI)	3.2 (2.8-3.7)	4.4 (1.8-5.65)		5.1 (2.4-7.1)		4.9 (2.5-6.6)		4.2 (1.9-4.7)	
	RR (95% CI)	N/A (Reference group)	**1.62 (1.20-2.17)**	**P=0.001**	**1.94 (1.48-2.54)**	**P<0.001**	**1.88 (1.40-2.54)**	**P<0.001**	1.51 (0.79-2.87)	P=0.20
	Frequency (%)	57/153 (37%)	62/132 (47%)		75/210 (36%)		41/105 (39%)		5/13 (38%)	

Exposure to tobacco smoke	RR (95% CI)	N/A (Reference group)	1.49 (0.93-2.39)	P=0.09	0.93 (0.61-1.44)	P=0.76	1.08 (0.65-1.79)	P=0.77	1.05 (0.32-3.37)	P=0.93
Pet ownership	Frequency (%)	10/153 (7%)	10/132 (8%)		15/210 (7%)		15/105 (14%)		1/13 (8%)	
	RR (95% CI)	N/A (Reference group)	1.06 (0.43-2.58)	P=0.90	0.99 (0.44-2.23)	P=0.99	2.15 (0.95-4.89)	P=0.07	0.006 (0.0001-2278)	P=0.68
Family history of asthma	Frequency (%)	35/153 (23%)	37/132 (28%)		78/210 (37%)		31/105 (30%)		3/13 (23%)	
	RR (95% CI)	N/A (Reference group)	1.13 (0.88-1.45)	P=0.32	**1.37 (1.11 – 1.70)**	**P=0.004**	1.17 (0.90-1.52)	P=0.23	1.01 (0.54 – 1.86)	P=0.98
Current eczema	Frequency (%)	8/153 (5%)	8/132 (6%)		17/210 (9%)		3/105 (3%)		1/13 (8%)	
	RR (95% CI)	N/A (Reference group)	1.17 (0.43-3.21)	P=0.76	1.59 (0.67-3.80)	P=0.29	0.53 (0.13-2.06)	P=0.36	1.51 (0.17-13.11)	P=0.71
Current rhinitis	Frequency (%)	42/153 (27%)	51/132 (39%)		147/210 (70%)		56/105 (53%)		6/13 (46%)	
	RR (95% CI)	N/A (Reference group)	**1.66 (1.01-2.74)**	**P=0.04**	**6.17 (3.89-9.78)**	**P<0.001**	**3.02 (1.79-5.09)**	**P<0.001**	2.26 (0.72-7.13)	P=0.17

4.6 Discussion

Our integrative approach of blending the data and bio-statistical expertise with clinical expert domain knowledge identified a framework for the discovery of stable and clinically meaningful asthma subtypes. Using two common clustering approaches (clustering after dimensionality reduction, and using all available variables) resulted in different clusters, which were not stable. We identified four features of asthma which exemplified the differences and similarities between the clusters in our initial analyses: age of onset, allergic sensitization, asthma severity and recent exacerbations. When we re-clustered the data using these four features, the cluster stability dramatically increased, and the analysis identified five clinically meaningful asthma subtypes (early-onset mild atopic asthma, early-onset mild non-atopic asthma, late-onset asthma, difficult asthma and exacerbation-prone asthma).

4.6.1 Limitations/strengths

One limitation of the clustering methodologies (including our analyses) is that for the selection of variables, the investigators rely on the data which is available. The majority of previous studies used similar data sources (e.g. detailed questionnaire responses, sensitization and lung function), but the variable choice for input into the model has varied.[18] We relied on a detailed clinical assessment carried out in our study. However, we cannot exclude the possibility that some potentially important variables were not collected.

Another limitation is that our study is cross-sectional, and precise information about the time dimension (particularly in relation to the age of onset of asthma) may be unreliable. However, cross-sectional datasets are ideal settings for data exploration and finding latent patterns. We could test various methodologies to ascertain the most robust one for our dataset. We acknowledge that adding more accurate information on onset and remission of symptoms to account for longitudinal changes could further improve asthma classification.

The strengths of our study include large number of phenotypically well-defined patients across the spectrum of asthma severity (from mild to severe), which improves generalizability.

Furthermore, to our knowledge, this is the first unsupervised analysis among children from a developing country, which offers a unique perspective on asthma subtypes in a population with different environmental exposures (and likely different genetic susceptibility) compared to studies in developed countries.

4.6.2 Interpretation

Data-driven methods have been used in both case/patient[18] and birth cohort studies,[16] and are invaluable tools for discovering complex patterns and structures in datasets. However, there has been little consistency in the results between different studies and no unified methodology, leading to a degree of scepticism in the clinical community about the value of these techniques.[5,23]

PCA has been used as both a standalone analysis,[14,36-38] and a data reduction technique prior to clustering.[12,21,24,39] Results from our PCA are consistent with previous studies in children, showing diversification with respect to lung function, demographics, medication use, symptom burden, and environmental factors.[11,21] One of the benefits of PCA is the reduction in dimensionality, which allows the description of the complex data using a smaller number of uncorrelated variables, while retaining as much information as possible. However, in our dataset, PCA has not substantially reduced dimensionality (from a total of 47 variables we identified 19 components with eigenvalue>1, which suggests that most variables may have been informative about different disease domains). PCA can be viewed as a method which separates signal and noise: the first dimensions extract the essential information, while the last ones are restricted to noise.[40] Intuitively, the reduction in noise should create more stable clusters; however, in the current study, inputting the principal components into the HC model yielded unstable clusters, which suggests that PCA did not differentiate between informative and non-informative variables. This could be a reflection of the dataset or the inherent heterogeneity of the disease.

It is generally considered that there is a linear relationship between the number of variables and stability of the model.[4] However, the clusters which emerged in the HC based on all available variables remained unstable, suggesting that it may not be useful to input all variables into the clustering algorithm, as the overloaded model may not be fully informative. Increasing the number of input variables increases the odds of the variables no longer being dissimilar (a feature important in differentiating clusters).[18] This introduces high degrees of collinearity among the variables, making it more difficult for the model to identify unique features, and some domains may be over-represented.

One of areas that remains to be addressed in statistical research is how to identify a meaningful set of features for cluster analysis using an unsupervised approach. In this study, we found that HC on PCA and HC on the raw data were less stable than HC on four selected features. This could be an artefact of the heterogeneity in the number of features. However, having a more meaningful semi-automated approach to feature selection for clustering is an area of machine learning research which may have a considerable impact on understanding disease heterogeneity.

In our study, by utilizing four informative features/domains of the disease which were identified by clinical experts who interpreted the results of the unsupervised analyses markedly increased cluster stability, and the results in clinical terms appeared much more meaningful. It is likely that these domains provide important information about asthma heterogeneity, which may be lost in the noise when using all collected variables or principal components. This may be analogous to our previous findings in a population-based birth cohort, in which dimensionality reduction suggested that out of >100 item responses in validated questionnaires, only 28 were informative for the discovery of disease subtypes.[14] Questions used to determine the presence of disease in most epidemiological studies (current wheezing, and wheezing apart from colds) were found to be redundant for understanding disease heterogeneity. This does not mean that these questions are not informative; they are key for ascertaining the presence of asthma syndrome, but are not informative when trying to uncover asthma subtypes. Thus, different domains of the disease may be required to identify disease subtypes than those used to diagnose asthma, or assess the control or response to treatment. Our results are consistent with the findings from the Childhood Asthma Management Program (which did not include children with severe asthma), which has reported that reproducible clusters with distinct clinical trajectories and different response to anti-inflammatory medications could be differentiated based on three groups of features (atopic burden, degree of airway obstruction, and history of exacerbation).[41]

In our study, severity was one of the key features for disaggregating the asthma syndrome, but there were children with moderate/severe asthma in each of the clusters. In the US Severe Asthma Research Program (SARP), a similar HC method was used to identify four subtypes of severe asthma in childhood, differing in age of onset, lung function, FeNO and medication use, but with an even distribution of severity among the clusters.[11] The Trousseau Asthma Program (TAP) identified a neutrophilic-driven severe asthma cluster that seemed to be resistant to corticosteroids.[21] In all three studies, severe asthma was not identified as an independent cluster. Rather, severe asthmatics were present in all clusters; in TAP, the proportion of severe asthmatics ranged from 5-10% across the clusters,[21] in SARP from 61-84% based on ATS criteria and 4-16% according to GINA,[11] and in our study the occurrence of moderate/severe asthmatics ranged from 8% in Cluster 3 to 65% in Cluster 1. The results from the current and other studies suggest that severe asthma is not a single entity, but rather the extreme end of spectrum of several different asthma endotypes.

Our study identified an exacerbation-prone cluster, which may be a separate endotype with unique underlying aetiology. A severe exacerbation cluster (which was predominantly allergy driven) was also described in the TAP cohort.[21] Recent analysis amongst SARP participants (both adults and children) has suggested that exacerbation-prone asthma may indeed be a distinct susceptibility phenotype, with implications for the targeting of exacerbation prevention strategies.[42] Exacerbation-prone asthma is not characterized only by asthma severity or control, and among SARP participants and in our study, a proportion of patients with exacerbation-prone asthma had non-severe asthma and normal lung function.[42]

The age at which a child initially wheezes has been described as a key discriminator of childhood wheeze phenotypes in multiple birth cohort studies, and our results which identified an early-onset and a late-onset asthma subtype are consistent with other previously published work.[7,15,19] However, unlike most previous studies, we identified both an early-onset non-atopic subtype and an early-onset atopic subtype.

Varying definition of allergic sensitization resulted in no material changes in our results. Using a model-based cluster analysis, Simpson *et al* have shown that sensitization comprises several different subtypes, each with unique association to asthma presence and severity,[43] and this finding was confirmed in another birth cohort.[44] For the prediction of future development of asthma, or asthma severity among patients with established disease, subtyping of sensitization may be crucially important.[8,43-45] However, our current analysis suggests that for the purpose of asthma subtyping, a simple definition of allergic sensitization would likely suffice.

In our study, most children with asthma had normal lung function. Although lung function was significantly diminished among children in the "Difficult asthma" cluster, most patients in this cluster had normal lung function, which is consistent with other populations.[46] Our analysis suggests that lung function may be less important for subtyping asthma, despite its perceived clinical importance for diagnosing and managing the disease. Our data also indicate that phenotyping asthma based on a single dimension of the disease (e.g. "eosinophilic" vs. "neutrophilic") is unlikely to be fully informative in the search for endotypes, or for precise treatment stratification. Blood eosinophilia was a significant feature of "Difficult", "Early-onset mild atopic" and "Late-onset asthma" clusters, suggesting that there are important shared mechanisms across different asthma subtypes.[8] Thus, while by definition each asthma endotype has a unique component in its pathophysiology,[1,2] these data indicate that some important mechanisms (e.g. T2-high) overlap between most endotypes.[5,8] This may also be reflected in the responses to treatment, and patients across different endotypes may display a spectrum of responses to therapies which target shared mechanisms.[5,41]

In conclusion, we identified four key features of asthma (age of onset, allergic sensitization, severity and exacerbations in the previous year), which may be informative for ascertaining asthma subtypes. This could represent a potential future framework to facilitate the discovery of endotypes in childhood asthma. Our results highlight that factors which are key determinants of asthma presence, severity or control may not be the most informative for determining disease endotypes.

4.7 References

1. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet* 2008; **372**(9643): 1107-19.

2. Lotvall J, Akdis CA, Bacharier LB, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol* 2011; **127**(2): 355-60.

3. A plea to abandon asthma as a disease concept. *Lancet* 2006; **368**(9537): 705.

4. Hennig C. Cluster-wise assessment of cluster stability: Department of Statistical Science, University College London, UK, 2006.

5. Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, Custovic A. Disaggregating asthma: Big investigation versus big data. *J Allergy Clin Immunol* 2017; **139**(2): 400-7.

6. Belgrave D, Simpson A, Custovic A. Challenges in interpreting wheeze phenotypes: the clinical implications of statistical learning techniques. *Am J Respir Crit Care Med* 2014; **189**(2): 121-3.

7. Belgrave DC, Custovic A, Simpson A. Characterizing wheeze phenotypes to identify endotypes of childhood asthma, and the implications for future management. *Expert review of clinical immunology* 2013; **9**(10): 921-36.

8. Custovic A, Sonntag HJ, Buchan IE, Belgrave D, Simpson A, Prosperi MC. Evolution pathways of IgE responses to grass and mite allergens throughout childhood. *J Allergy Clin Immunol* 2015; **136**(6): 1645-52 e1-8.

9. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 2012; **18**(5): 716-25.

10. Custovic A, Ainsworth J, Arshad H, et al. The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together. *Thorax* 2015; **70**(8): 799-801.

11. Fitzpatrick AM, Teague WG, Meyers DA, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *J Allergy Clin Immunol* 2011; **127**(2): 382-9 e1-13.

12. Haldar P, Pavord ID, Shaw DE, et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008; **178**(3): 218-24.

13. 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**(4): 315-23.

14. Smith JA, Drake R, Simpson A, Woodcock A, Pickles A, Custovic A. Dimensions of respiratory symptoms in preschool children: population-based birth cohort study. *Am J Respir Crit Care Med* 2008; **177**(12): 1358-63.

15. Henderson J, Granell R, Heron J, 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.

16. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. *Curr Allergy Asthma Rep* 2015; **15**(7): 38.

17. Deliu M, Belgrave D, Sperrin M, Buchan I, Custovic A. Asthma phenotypes in childhood. *Expert review of clinical immunology* 2016: 1-9.

18. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of Asthma Subtypes Using Clustering Methodologies. *Pulm Ther* 2016; **2**: 19-41.
19. Belgrave DC, Simpson A, Semic-Jusufagic A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. *J Allergy Clin Immunol* 2013; **132**(3): 575-83.e12.
20. Patrawalla P, Kazeros A, Rogers L, et al. Application of the asthma phenotype algorithm from the Severe Asthma Research Program to an urban population. *PLoS One* 2012; **7**(9): e44540.
21. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. *Eur Respir J* 2012; **40**(1): 55-60.
22. Benton AS, Wang Z, Lerner J, Foerster M, Teach SJ, Freishtat RJ. Overcoming heterogeneity in pediatric asthma: tobacco smoke and asthma characteristics within phenotypic clusters in an African American cohort. *J Asthma* 2010; **47**(7): 728-34.
23. Belgrave D, Custovic A. The importance of being earnest in epidemiology. *Acta Paediatr* 2016; **105**(12): 1384-6.
24. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *Am J Respir Crit Care Med* 2013; **188**(11): 1303-12.
25. Sackesen C, Karaaslan C, Keskin O, et al. The effect of polymorphisms at the CD14 promoter and the TLR4 gene on asthma phenotypes in Turkish children with asthma. *Allergy* 2005; **60**(12): 1485-92.
26. Sahiner UM, Semic-Jusufagic A, Curtin JA, et al. Polymorphisms of endotoxin pathway and endotoxin exposure: in vitro IgE synthesis and replication in a birth cohort. *Allergy* 2014; **69**(12): 1648-58.
27. Crapo RO, Casaburi R, Coates AL, et al. Guidelines for methacholine and exercise challenge testing-1999. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. *American journal of respiratory and critical care medicine* 2000; **161**(1): 309-29.
28. Crapo RO, Casaburi R, Coates AL, et al. Guidelines for methacholine and exercise challenge testing-1999. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. *American journal of respiratory and critical care medicine* 2000; **161**(1): 309-29.
29. Popa V. ATS guidelines for methacholine and exercise challenge testing. *American journal of respiratory and critical care medicine* 2001; **163**(1): 292-3.
30. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *European respiratory journal* 2005; **26**(2): 319-38.
31. Beydon N, Davis SD, Lombardi E, et al. An official American Thoracic Society/European Respiratory Society statement: pulmonary function testing in preschool children. *American journal of respiratory and critical care medicine* 2007; **175**(12): 1304-45.
32. Stanojevic S, Wade A, Cole TJ, et al. Spirometry centile charts for young Caucasian children: the Asthma UK Collaborative Initiative. *Am J Respir Crit Care Med* 2009; **180**(6): 547-52.

33. Quanjer PH, Stanojevic S, Cole TJ, et al. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *European Respiratory Journal* 2012; **40**(6): 1324-43.
34. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
35. C H. Cluster-wise assessment of cluster stability. London UK: University College London, 2006.
36. Rodriguez A, Vaca M, Oviedo G, et al. Urbanisation is associated with prevalence of childhood asthma in diverse, small rural communities in Ecuador. *Thorax* 2011; **66**(12): 1043-50.
37. Chawes BL, Stokholm J, Bonnelykke K, Brix S, Bisgaard H. Neonates with reduced neonatal lung function have systemic low-grade inflammation. *J Allergy Clin Immunol* 2015; **135**(6): 1450-6 e1.
38. Clemmer GL, Wu AC, Rosner B, et al. Measuring the corticosteroid responsiveness endophenotype in asthmatic patients. *J Allergy Clin Immunol* 2015; **136**(2): 274-81 e8.
39. Weatherall M, Travers J, Shirtcliffe PM, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 2009; **34**(4): 812-8.
40. Husson F JJ, Pages J. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? Agrocampus Ouest, FR: Agrocampus Ouest, 2010.
41. Howrylak JA, Fuhlbrigge AL, Strunk RC, et al. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. *J Allergy Clin Immunol* 2014; **133**(5): 1289-300, 300 e1-12.
42. Denlinger LC, Phillips BR, Ramratnam S, et al. Inflammatory and Comorbid Features of Patients with Severe Asthma and Frequent Exacerbations. *Am J Respir Crit Care Med* 2017; **195**(3): 302-13.
43. Simpson A, Tan VY, Winn J, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med* 2010; **181**(11): 1200-6.
44. Lazic N, Roberts G, Custovic A, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* 2013; **68**(6): 764-70.
45. Simpson A, Lazic N, Belgrave DC, et al. Patterns of IgE responses to multiple allergen components and clinical symptoms at age 11 years. *J Allergy Clin Immunol* 2015; **136**(5): 1224-31.
46. Bush A, Saglani S. Management of severe asthma in children. *Lancet* 2010; **376**(9743): 814-25.

4.7 Supplementary Material/ Appendix

4.7.1 Methods

Statistical Methods

Variables used

1) Binary variables
    a. Interview-derived: sex, physician diagnosed allergic rhinitis, allergic conjunctivitis, eczema, family history of asthma, exposure to tobacco, pet ownership
    b. Medications: use of Long-acting β_2- agonist, use of montelukast, inhaled corticosteroid dose (0, less than 400, greater than 400 *BDP equivalent to beclomethasone), use of short acting β_2 - agonist
    c. Atopy: wheel 3mm greater than negative control to at least one of: cat, dog, cockroach, tree, weed, grass, house dust mite, moulds, serum total IgE
    d. Eosinophil % (0-0.15, 0.15-0.30, 0.3-0.5, >0.5)
2) Count variables
    a. Number of asthma exacerbations within the last 12 months, number of hospitalizations for asthma ever
3) Continuous variables
    a. Age, age of asthma onset, BMI (standardised for age, growth, sex)
    b. Lung function:
        i. FEV_1 % predicted, FVC % predicted, FEV_1/FVC, FEF_{25-75}
        ii. Bronchodilator reversibility: greater than or equal to 12% increase in FEV_1 following administration of 200μg inhaled albuterol

iii. Airway hyperresponsiveness: concentration of methacholine required to produce a 20% decline in FEV_1 (PC_{20}) less than or equal to 8mg/ml[E1] or greater than or equal to 10% reduction in FEV_1 following exercise challenge[E2].

Variables were used in their raw format apart from: (1) 'inhaled corticosteroid use' which was categorised with equal frequency binning and projected into a dummy variable (ICS =0, ICS <400mg, ICS>400 BDP equivalent); (2) 'blood eosinophil count' (<0.15, 0.15-0.30, 0.30-0.5, >0.5); (3) 'asthma severity (mild, moderate, severe); (4) 'asthma attacks' (0,1,≥2); (5) 'age of onset'(<5, 5-11, >11); and (6) 'asthma hospitalizations' (<2, ≥2). Our dummy variables were created using clinically meaningful categories, except for age of onset which was arbitrarily chosen. Dummy variables were only used in model 1.

Continuous variables were transformed into z-scores to simplify interpretation (whereby coefficients refer to a change of 1 standard deviation), and remove skew. No missing values were present apart from 'methacholine challenge' and 'exercise challenge test' (which were only available for approximately half of the children). These two variables were not included in the analysis.

Model 1: Hierarchical clustering after dimensionality reduction

Principal component analysis was performed on 45 variables in the dataset (which included aforementioned dummy variables). Variables with loadings above 0.3 on a particular principal component were chosen to represent that component. The loadings threshold was set lower than what is seen in most literature to take into account the large amount of binary data, as these naturally load lower. Orthogonal/varimax rotation was performed.

Hierarchical clustering using Ward's method and Euclidean distance was performed on principal components. As sensitivity analysis, other distance measures (such as Minkowski, etc.) and methods (single-link, centroid, average) were also tested.

Model 2: Hierarchical clustering using raw data

In Model 2, dummy variables (except inhaled corticosteroid use) were removed leaving 38 raw variables. Hierarchical clustering was then performed and results were compared to the clusters identified in Model 1.

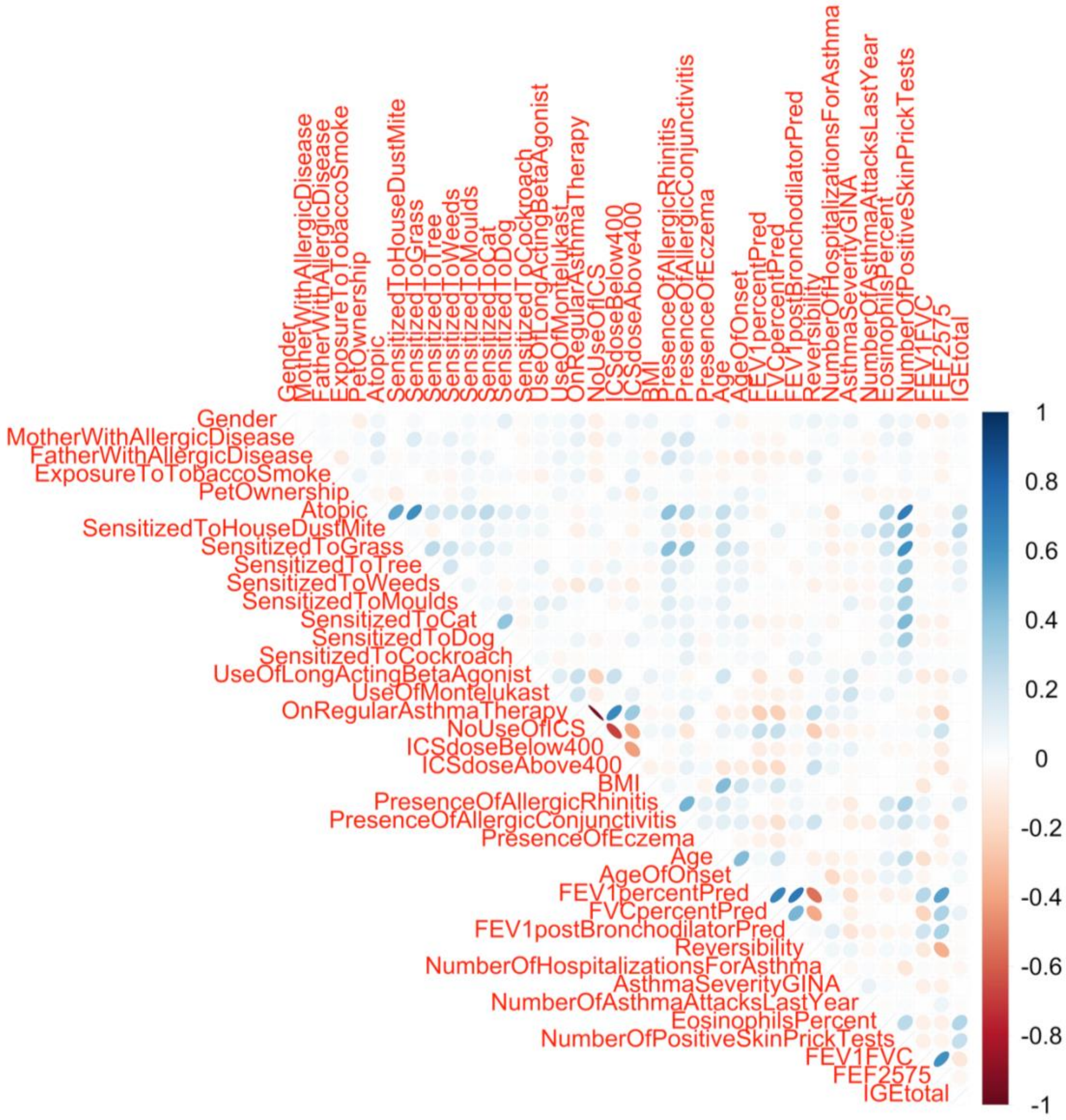
Model 3: Identification of key stable/important features and re-clustering

Clusters in Models 1 and 2 were generally not stable. After comparing the results, four key features were identified: age of onset, asthma severity, exacerbations, atopy. We performed a series of hierarchical cluster analyses on these four features, each time varying measures of atopy: atopic, polysensitised, monosensitized, total number of positive skin prick tests, total IgE. Specifics of atopic variables:

- *Atopic:* binarised variable, atopic (positive skin prick test to any allergen) yes/no
- *Sensitization:* categorical variable (0, 1, 2), 0=non atopic, 1=monosensitized (positive skin prick test to only one allergen), 2=polysensitized (positive skin prick test to more than 1 allergen)
- *Total IgE:* quantitative variable for total IgE levels in blood (scaled)

4.7.2 Results

Figure E4.1: *Correlation matrix of dataset*



BMI = Body mass index, FEF = forced expiratory flow, ICS = inhaled corticosteroids, SPT = skin prick test. Large blue circles show high positive correlation while red circles show negative correlation. Empty boxes denote a non-significant correlation.

Table E4.1: *Variables in the dataset used for analyses.* Variables in dataset but not directly used for analysis: methacholine challenge, exercise induced fall in FEV1, number of eosinophils, total number of positive skin prick tests.*

Variables in HC on PCA (Model 1)	**Variables in HC on raw data (Model 2)**
Sex	Sex
Mother with allergic disease	Mother with allergic disease
Father with allergic disease	Father with allergic disease
Exposure to tobacco smoke	Exposure to tobacco smoke
Pet ownership	Pet ownership
Atopic	Atopic
Sensitized to house dust mite	Sensitized to house dust mite
Sensitized to grass	Sensitized to grass
Sensitized to trees	Sensitized to trees
Sensitized to weeds	Sensitized to weeds
Sensitized to moulds	Sensitized to moulds
Sensitized to cat	Sensitized to cat
Sensitized to dog	Sensitized to dog
Sensitized to cockroach	Sensitized to cockroach
Blood Eosinophil % 0.15-0.3	Blood Eosinophil %
Blood Eosinophil % 0.3-0.5	IgE total
Blood Eosinophil % >0.50	Use of long-acting beta$_2$ agonist
IgE total	Use of Montelukast
Use of long-acting beta$_2$ agonist	Use of regular controller medication
Use of Montelukast	No use of inhaled corticosteroids
Use of regular controller medication	Inhaled corticosteroid dose <400
No use of inhaled corticosteroids	Inhaled corticosteroid dose >400
Inhaled corticosteroid dose <400	BMI
Inhaled corticosteroid dose >400	Presence of allergic rhinitis
BMI	Presence of allergic conjunctivitis
Presence of allergic rhinitis	Presence of eczema

Presence of allergic conjunctivitis	Age at follow-up
Presence of eczema	FEV_1 % predicted
Age of onset below 5 years	FEV_1/FVC % predicted
Age of onset 5-11 years	FEF 25-75
Age of onset above 11 years	FEV_1 post bronchodilator % predicted
FEV_1 % predicted	Reversibility
FVC % predicted	FEV_1 % predicted
FEV_1/FVC % predicted	Number of hospitalizations for asthma ever
FEF 25-75	Number of asthma attacks within last year
FEV_1 post bronchodilator % predicted	Mild asthma
Reversibility	Moderate-severe asthma
Less than 2 hospitalizations for asthma ever	Severe asthma
2 or more hospitalizations for asthma ever	
No asthma attacks within last year	
1 asthma attack within last year	
2 or more asthma attacks within last year	
Mild asthma	
Moderate-severe asthma	
Severe asthma	

Table E4.2: *Eigenvalue and variance of the first 19 components with eigenvalues >1*

	Eigenvalue	**Percentage of variance**	**Cumulative percentage of variance**
Component 1	4.29	8.95	8.95
Component 2	3.19	6.66	15.61
Component 3	2.56	5.34	20.96
Component 4	2.34	4.88	25.84
Component 5	2.11	4.40	30.24
Component 6	2.09	4.37	34.62
Component 7	1.93	4.04	38.66
Component 8	1.86	3.89	42.55
Component 9	1.80	3.76	46.32
Component 10	1.69	3.52	49.84
Component 11	1.49	3.11	52.96
Component 12	1.39	2.91	55.87
Component 13	1.38	2.87	58.75
Component 14	1.21	2.53	61.29
Component 15	1.17	2.45	63.74
Component 16	1.15	2.41	66.16
Component 17	1.10	2.30	68.46
Component 18	1.07	2.23	70.69
Component 19	1.05	2.19	72.89

Table E4.3: *Variable loadings on first 5 components*

a) Principal components 1 and 2

Dimension 1		**Dimension 2**	
Variable	**Loading**	**Variable**	**Loading**
Use of regular controller medication	0.66	Atopic	0.71
Moderate-severe asthma	0.55	Sensitized to grass	0.65
Reversibility	0.48	Presence of allergic rhinitis	0.55
Use of long-acting beta$_2$ agonist	0.43	Age at follow-up	0.51
ICS dose <400	0.39	Presence of allergic conjunctivitis	0.47
ICS dose > 400	0.33	Less than 2 hospitalizations ever	0.32
1 asthma attack within last year	0.32	Sensitized to tree	0.32
Use of Montelukast	0.3	Sensitized to cat	0.31
FEV$_1$/FVC	-0.33	Age of onset above 11 years	0.3
FEV$_1$post bronchodilator % predicted	-0.47	More than 2 hospitalizations for asthma ever	-0.32
FVC % predicted	-0.54	Age of onset below 5 years	-0.36
FEF2575	-0.58		
Mild asthma	-0.62		
No use of ICS	-0.66		
FEV$_1$%pred	-0.72		

ICS: inhaled corticosteroid

b) Principal components 3 and 4

Dimension 3		Dimension 4	
Variable	**Loading**	**Variable**	**Loading**
Mild asthma	0.64	More than 2 hospitalizations for asthma ever	0.49
Presence of allergic conjunctivitis	0.44	Male	0.47
Use of regular controller medication	0.39	FEV_1%pred	0.47
1 asthma attack within last year	0.36	FVC% pred	0.42
Reversibility	0.35	FEV_1 post bronchodilator % pred	0.4
FEV_1 post bronchodilator % pred	0.34	Blood Eo >0.5%	-0.3
Use of long-acting beta$_2$ agonist	-0.39	Female	-0.47
No use of ICS	-0.4	Less than 2 hospitalizations for asthma ever	-0.49
Moderate-severe asthma	-0.62		

c) Principal component 5

Dimension 5	
Variable	**Loading**
No use of ICS	0.41
More than 2 hospitalizations for asthma ever	0.33
Presence of allergic rhinitis	0.31
Atopic	0.31
Blood Eo >0.50%	-0.3
BMI	-0.3
Less than 2 hospitalizations for asthma ever	-0.33
Use of regular controller medication	-0.4
ICS dose < 400	-0.48

Table E4.4: *Characteristics of each cluster for model 1 (HC after PCA dimensionality reduction).Quantitative variables are represented as mean (standard deviation and interquartile range). Binary and categorical variables are represented as proportions (%). Binary and categorical variables are calculated using chi-squared test of significance; continuous variables are calculated using Kruskal-Wallis test of significance.*

Variable		**Cluster 1** **n = 102**	**Cluster 2** **n = 70**	**Cluster 3** **n = 117**	**Cluster 4** **n = 149**	**Cluster 5** **n = 175**	**p-value**
Sex	Male	61/102, 59%	56/70, 80%	84/117, 72%	72/149, 48%	95/175, 54%	<0.001
	Female	41/102, 41%	14/70, 20%	33/117, 28%	77/149, 52%	81/175, 46%	<0.001
Mother with allergic disease		8/102, 7%	19/70, 28%	36/117, 31%	30/149, 20%	24/175, 15%	0.002
Father with allergic disease		6/102, 5%	40/70, 57%	35/117, 30%	20/149, 13%	19/175, 11%	0.01
Exposure to tobacco smoke		28/102, 27%	52/70, 74%	52/117, 44%	60/149, 40%	48/175, 27%	0.15
Pet ownership		7/102, 6%	37/70, 53%	16/117, 14%	23/149, 15%	12/175, 7%	0.28
Atopic		75/102, 75%	31/70, 48%	105/117, 89%	49/149, 32%	102/175, 58%	<0.001
Sensitized to house dust mite		45/102, 41%	7/70, 10%	25/117, 21%	22/149, 15%	65/175, 37%	<0.001
Sensitized to grass		30/102, 9%	17/70, 25%	90/117, 77%	19/149, 13%	56/175, 32%	<0.001
Sensitized to trees		7/102, 6%	0/70	15/117, 13%	2/149, 1%	5/175, 3%	<0.001
Sensitized to weeds		5/102, 5%	0/70	7/117, 6%	0/149	14/175, 8%	0.008

Sensitized to moulds	8/102, 7%	2/70, 3%	12/117, 10%	4/149, 3%	12/175, 7%	0.03
Sensitized to cat	11/102, 10%	0/70	21/117, 18%	2/149, 1%	15/175, 9%	<0.001
Sensitized to dog	3/102, 3%	1/70, 1%	9/117, 8%	0/149	2/175, 1%	<0.001
Sensitized to cockroach	2/102, 2%	0/70	3/117, 3%	1/149, 1%	5/175, 3%	0.6
Blood Eosinophil % 0.15-0.3	8/102, 8%	9/70, 13%	0/117	1/149, 1%	0/175	0.5
Blood Eosinophil % 0.3-0.5	17/102, 16%	16/70, 23%	8/117, 7%	10/149, 7%	4/175, 2%	0.006
Blood Eosinophil % >0.50	83/102, 81%	44/70, 64%	109/117, 93%	130/149, 87%	173/175, 99%	<0.001
IgE total	365.5 (534.9, 34-440.3)	79.1 (115.5, 11.5-83.5)	293.8 (582, 49-272)	108.8 (276.1,12-106)	233.1 (476,42-221.5)	<0.001
Use of long-acting beta$_2$ agonist	43/102, 41%	4/70, 6%	3/117, 3%	0/149	0/175	<0.001
Use of Montelukast	24/102, 23%	10/70, 14%	8/117, 7%	2/149, 0.01%	7/175, 4%	<0.001

Use of regular controller medication	76/102, 72%	61/70, 88%	105/117, 89%	133/149, 90%	10/175, 6%	<0.001
No use of inhaled corticosteroids	30/102, 29%	8/70, 12%	10/117, 9%	15/149, 10%	174/175, 99%	<0.001
Inhaled corticosteroid dose <400	55/102, 52%	44/70, 64%	74/117, 63%	83/149, 56%	6/175, 3%	<0.001
Inhaled corticosteroid dose >400	22/102, 22%	10/70, 14%	33/117, 28%	49/149, 33%	1/175, 0.01%	<0.001
BMI	18.6 (3.6, 16-22.2)	19.2 (3.9, 16.3-22.2)	16.9 (3.1, 15.8-19.1)	17.5 (3.4, 15.4-19.2)	17.9 (3.4, 15.9-21.1)	0.08
Presence of allergic rhinitis	50/102, 50%	15/70, 21%	92/117, 78%	35/149, 24%	110/175, 63%	<0.001
Presence of allergic conjunctivitis	6/102, 6%	6/70, 9%	71/117, 61%	8/149, 5%	21/175, 12%	<0.001
Presence of eczema	9/102, 9%	3/70, 4%	11/117, 9%	6/149, 4%	8/175, 5%	0.35
Age of onset below 5 years	39/102, 39%	19/70, 19%	25/117, 21%	88/149, 61%	64/175, 36%	<0.001
Age of onset 5-11 years	38/102, 37%	56/70, 81%	30/117, 26%	56/149, 38%	95/175, 54%	0.02

Age of onset above 11 years	25/102, 24%	1/70, 1%	62/117, 53%	3/149, 2%	16/175, 9%	0.02
FEV_1 % predicted	75.6 (10.9, 70.25-82)	93.8 (11.7, 86-102)	82 (11.8, 74-90)	83.8 (11.2, 77-90)	96.7 (11.3, 90-104)	<0.001
FVC % predicted	89.9 (14.4,81-99.8)	100.4 (12.7, 91-111.8)	90.4 (16.9, 81-102)	92 (11.4, 84-99)	105.1 (11.3, 97-112)	<0.001
FEV_1/FVC % predicted	80.8 (8.8, 74.3-86.8)	86.7 (6.0, 78.7-90.3)	86.2 (5.6, 78.6-90.2)	86.8 (5.9, 85.7-91.2)	87.6 (6.8, 85-92)	<0.001
FEF 25-75	88.3 (11.2, 83-96)	82 (21.9, 76-88)	72.4 (17.7, 61-84)	84.6 (22.9, 77-89.1)	97.5 (6.8, 80-111)	<0.001
FEV_1 post bronchodilator % predicted	65.3(22.8, 50-75.5)	108.1 (12.4, 101-115)	100.2 (10.7, 99.7-105.2)	100.3 (11.9, 99.8-108.2)	105.2 (12.6, 97-113.3)	<0.001
Reversibility	17.9 (10.1, 8.1-18.6)	18.7 (10.3, 9.1-19.6)	24.4 (18.1, 13.1-29.3)	18.1 (13.7, 8.2-21.8)	9.8 (12.6, 0-14.4)	<0.001
Less than 2 hospitalizations for asthma ever	104/102, 99%	40/70, 58%	0/117	0/149	171/175, 98%	<0.001
2 or more hospitalizations for asthma ever	1/102, 1%	29/70, 42%	0/117	0/149	1/175, 0.6%	<0.001
No asthma attacks within last year	54/102, 53%	29/70, 42%	71/117, 61%	96/149, 64%	140/175, 80%	<0.001
1 asthma attack within last year	15/102, 15%	14/70, 20%	42/117, 36%	37/149, 25%	13/175, 7%	<0.001

2 or more asthma attacks within last year	23/102, 22%	27/70, 39%	4/117, 3%	16/149, 11%	22/175, 13%	<0.001
Mild asthma	2/102, 2%	59/70, 85%	103/117, 88%	142/149, 97%	174/175, 99%	<0.001
Moderate-severe asthma	95/102, 90%	10/70, 14%	14/117, 11%	6/149, 3%	1/175, 1%	<0.001
Severe asthma	8/102, 8%	1/70, 1%	1/117, 1%	1/149, 1%	0/175	<0.001

***Table E4.5**: Characteristics of each cluster for model 2 (HC using all available variables).Quantitative variables are represented as mean (standard deviation and interquartile range). Binary and categorical variables are represented as proportions (%). Binary and categorical variables are calculated using chi-squared test of significance; continuous variables are calculated using Kruskal-Wallis test of significance.*

Variable		**Cluster 1** **n = 168**	**Cluster 2** **n = 100**	**Cluster 3** **n = 103**	**Cluster 4** **n = 223**	**Cluster 5** **n = 19**	**p-value**
Sex	Male	76/168, 55%	62/100, 62%	72/103, 70%	149/223, 67%	13/19, 68%	0.07
	Female	92/168, 45%	38/100, 38%	31/103, 30%	74/223, 33%	6/19, 32%	0.07
Mother with allergic disease		43/168, 26%	24/100, 24%	16/103, 16%	32/223, 14%	4/19, 21%	0.04
Father with allergic disease		28/168, 17%	12/100, 12%	17/103, 17%	24/223, 11%	2/19, 11%	0.40
Exposure to tobacco smoke		78/168, 46%	36/100, 36%	47/103, 46%	72/223, 32%	7/19, 37%	0.03
Pet ownership		10/168, 6%	13/100, 13%	9/103, 9%	18/223, 8%	1/19, 5%	0.41
Atopic		110/168, 65%	81/100, 81%	68/103, 67%	86/223, 39%	16/19, 84%	<0.001
Sensitized to house dust mite		52/168, 31%	39/100, 39%	31/103, 31%	42/223, 19%	13/19, 68%	<0.001
Sensitized to grass		64/168, 38%	58/100, 58%	44/103, 43%	52/223, 23%	9/19, 47%	<0.001
Sensitized to trees		9/168, 5%	7/100, 7%	5/103, 5%	5/223, 2%	2/19, 11%	0.81
Sensitized to weeds		8/168, 5%	3/100, 3%	3/103, 3%	12/223, 5%	1/19, 5%	0.20

Sensitized to moulds	14/168, 8%	6/100, 6%	5/103, 5%	12/223, 5%	1/19, 5%	0.75
Sensitized to cat	13/168, 8%	8/100, 8%	16/103, 16%	13/223, 5%	1/19, 5%	0.06
Sensitized to dog	5/168, 3%	1/100, 1%	6/103, 6%	4/223, 2%	0/19	0.17
Sensitized to cockroach	3/168, 2%	2/100, 2%	3/103, 3%	3/223, 1%	0/19	0.85
Blood Eosinophil %	2.0 (3.4, 0.7-5.8)	4.1 (2.4, 1.7-6.7)	2.8 (3.1, 1.1-6.1)	2.2 (2.1, 0.8-4.5)	6.1 (4.5, 2.1-15.7)	0.23
IgE total	136.1 (167.4, 24.8-172.5)	240.6 (296.2, 55.7-245)	250.5 (294.9, 31-345.5)	121.4 (153.4, 21-157.5)	2152.9 (1005.3, 1680-2169)	<0.001
Use of long-acting beta$_2$ agonist	19/168, 11%	8/100, 8%	12/103, 12%	7/223, 3%	5/19, 26%	<0.001
Use of Montelukast	20/168, 12%	7/100, 7%	9/103, 9%	14/223, 6%	1/19, 5%	0.33
Use of regular short acting beta$_2$ agonist	134/168, 80%	62/100, 62%	76/103, 76%	98/223, 44%	14/19, 74%	<0.001
No use of inhaled corticosteroids	34/168, 20%	41/100, 41%	27/103, 27%	130/223, 58%	5/19, 26%	<0.001
Inhaled corticosteroid dose <400	82/168, 49%	52/100, 52%	52/103, 50%	66/223, 30%	10/19, 53%	<0.001
Inhaled corticosteroid dose >400	51/168, 30%	7/100, 7%	24/103, 23%	27/223, 12%	4/19, 21%	<0.001

BMI	18.2 (3.8, 15.5-20.1)	18.3 (3.6, 15.5-20.9)	17.9 (3.5, 15.7-19.9)	18.8 (3.4, 16.1-21.3)	19.2 (3.6, 15.7-22.4)	<0.001
Presence of allergic rhinitis	88/168, 52%	69/100, 69%	50/103, 49%	84/223, 38%	11/19, 58%	<0.001
Presence of allergic conjunctivitis	39/168, 23%	34/100, 34%	20/103, 20%	17/223, 8%	2/19, 11%	<0.001
Presence of eczema	12/168, 7%	3/100, 3%	8/103, 8%	13/223, 6%	1/19, 5%	0.63
Age at follow-up	3.1 (2.1, 2.5-4.8)	11.2 (3, 7-10.7)	5.8 (2.1, 1.8-7.2)	9.2 (2.7, 7.2-10.7)	6.6 (2.2, 3.4-8.2)	<0.001
FEV_1 % predicted	79.92 (10.8, 74-87)	94.0 (10.7, 85.8-101.3)	72.6 (9.8, 68-78)	95.2 (9.5, 87-102)	91.2 (17.3, 80.5-99.5)	<0.001
FVC % predicted	85.6 (11.5, 78-93.25)	106.0 (14.6, 94.8-115.3)	89.1 (15.9, 81-100)	102.1 (9.5, 94.5-108)	102.2 (15.1, 94-109.5)	<0.001
FEV_1/FVC % predicted	88.9 (5.3, 85.8-92)	85.7 (6.1, 83-90)	78.1 (7.3, 73-83)	87.3 (5.3, 84-90)	84.6 (6.9, 81-89)	<0.001
FEF 25-75	81.8 (22.8, 68-90)	86.2 (24.2, 70-103)	54.8 (13.6, 45.5-61)	92.8 (21.4, 76.5-107)	82.5 (22.5, 69-89)	<0.001
FEV_1 post bronchodilator % predicted	92.8 (11.4, 86-100.8)	106.2 (10.8, 98.5-112)	94.2 (12.1, 86-101)	105.7 (11.5, 97-113)	103.6 (18.9, 92.3-105.8)	<0.001

Reversibility	16.6 (6.3, 12.9-20)	14.3 (7.1, 12.1-17.9)	31.6 (21.8, 14.1-43.2)	11.9 (6.8, 5.15-15.7)	14.6 (6.3, 12.5-15.8)	<0.001
Number of hospitalizations for asthma ever	0.8 (0.2, 0.4-2.2)	0.4 (0.2, 0-1.1)	0.5 (0.3, 0.1-1.9)	0.1 (0.2, 0-1.8)	0.1 (0.2, 0-1)	0.63
Number of asthma attacks within last year	0.8 (1.8, 0.4-1.7)	0.6 (2.3, 0-7)	0.7 (2.3, 0.2-1.3)	0.2 (1.7, 0.1-1.3)	2.5 (2.2, 0.9-4.8)	<0.001
Mild asthma	117/168, 69%	92/100, 92%	47/103, 46%	202/223, 90%	11/19, 58%	<0.001
Moderate-severe asthma	43/168, 26%	8/100, 8%	54/103, 52%	20/223, 9%	8/19, 42%	<0.001
Severe Asthma	8/168, 5%	0/100	2/103, 2%	1/223, 1%	0/19	0.01

Table E4.6: *Crosstabs of cluster subject allocation. Results are presented as proportions. The columns represent the cluster membership from the analysis done on principal components. Rows represent cluster membership from the analysis done on raw data. Highlighted values indicate highest overlap and likely corresponding clusters. Association was measured by the chi-squared test.*

		HC after PCA dimensionality reduction					
HC using all available variables		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
	Cluster 1	37 (22%) p=0.26	17 (10%) P<0.001	34 (20%) p=0.02	**65 (39%)** **p<0.001**	15 (9%) p<0.001	168
	Cluster 2	5 (5%), p=0.82	6 (6%) p=0.06	**45 (45%)** **p<0.001**	8 (8%) p<0.001	36 (36%) p<0.001	100
	Cluster 3	**45 (44%)** **p<0.001**	8 (8%) p=0.01	29 (28%) p=0.22	16 (15%) p=0.008	5 (5%) p<0.001	103
	Cluster 4	8 (4%), p0<0.001	37 (16%) p=0.002	6 (3%) p<0.001	58 (26%) p=0.51	**114 (51%)** **p<0.001**	223
	Cluster 5	7 (37%) p<0.001	**2 (11%)** **p<0.001**	3 (16%) p=0.001	2 (11%) p=0.04	5 (26%) p<0.001	19
Total		102 17%	70 11%	117 19%	149 24%	175 29%	**613** **100%**

Table E4.7: *Clinical comparison of outcomes based on the univariate analysis. Clinical comparison of outcomes based on the univariate analysis. The bolded parts indicate the major differences between the two results*

HC after PCA dimensionality reduction	HC using all available variables
Cluster 1 - *Moderate-severe asthma* - *Diminished lung function* - ***2 or more exacerbations, less than 2 hospitalizations*** - ***High IgE*** - High medication use - Mild eosinophilia	**Cluster 3** - *Moderate-severe asthma* - *Diminished lung function* - ***1 attack*** - ***Sensitized to cat, dog and tree*** - Reversible airways
Cluster 2 - *Frequent hospitalizations* - *Frequent attacks* - *Late-onset* - Moderate eosinophilia - Moderate medication use - Male - Family history - Normal lung function	**Cluster 5** - *Frequent attacks* - *Late-onset* - Atopic - High IgE - On LABA
Cluster 3 - *Late- onset* - *Multiple atopy* - *High BMI* - *Allergic rhinitis/conjunctivitis* - *Mild asthma* - 1 attack, less than 2 hospitalizations - ***Diminished lung function and high reversibility*** - Male - Normal eosinophils - Family history - Exposure to tobacco - Low medication use	**Cluster 2** - *Late- onset* - *Multiple atopy* - *High BMI* - *Allergic rhinitis/conjunctivitis* - *Mild asthma* - No attacks - ***Good lung function***
Cluster 4	**Cluster 1**

- *Early onset* - *Female* - *Non atopic* - *High steroid doses* - *Slightly diminished lung function* - ***Mild asthma*** - 1 attack per year	- *Early onset* - *Female* - *Slightly atopic; lower total IgE* - *High steroid and medication use* - *Diminished lung function* - ***Severe asthma*** - Family history - Exposed to tobacco
Cluster 5 - *Mild asthma* - *Good lung function* - *Low medication use* - ***Sensitized to HDM, weed*** - *No attacks* - *Less than 2 hospitalizations* - Hypereosinophilic	**Cluster 4** - *Mild asthma* - *Good lung function* - *Low medication use* - ***Non-atopic*** - ***Low IgE*** - Male

Table E4.8: *Cluster stability for the HC after PCA dimensionality reduction, HC using all available variables, and HC using the "informative" subset of features. The mean values for the bootstrapping samples indicating stability. Good stability is considered to have a bootstrap mean>0.75. Stable clusters highlighted in bold: HC on principal components producing only one stable cluster (Cluster 1), HC using all available data producing two stable clusters (Clusters 2 and 5), while in HC using the "informative" subset of features, all clusters were stable.*

HC after PCA dimensionality reduction	HC using all available variables	HC using the "informative" subset of features
Cluster 1: 0.98 Cluster 2: 0.23 Cluster 3: 0.41 Cluster 4: 0.60 Cluster 5: 0.19		
	Cluster 1: 0.59 **Cluster 2: 0.82** Cluster 3: 0.53 Cluster 4: 0.61 **Cluster 5: 0.86**	
		Cluster 1: 1.00 **Cluster 2: 0.99** **Cluster 3: 0.99** **Cluster 4: 1.00** **Cluster 5: 1.00**

Table E4.9: *Univariate logistic regression analysis using sensitization status as ordinal variable (non-atopic, monosensitized, polysensitized).*
**Coeff: The coefficient translates into a value of how likely a child is assigned to that cluster based on the variable response.*

Variable	**Cluster 1 (n=132)**		**Cluster 2 (n=135)**		**Cluster 3 (n=181)**		**Cluster 4 (n=151)**		**Cluster 5 (n=14)**	
	Coeff	p-value	Coeff	p-value	Coeff	p-value	Coeff	p-value	Coeff	p-value
Age of Onset	**-0.03**	**0.04**	**-0.15**	**<0.001**	**0.31**	**<0.001**	**-0.12**	**<0.001**	-0.006	0.25
Asthma attacks	0.0008	0.96	**-0.03**	**0.04**	**-0.05**	**0.009**	**-0.04**	**0.01**	**0.12**	**<0.001**
Sensitization status	-0.004	0.79	**0.15**	**0.003**	**0.10**	**<0.001**	**-0.26**	**<0.001**	0.004	0.45
Asthma Severity	**0.39**	**<0.001**	**-0.11**	**<0.001**	**-0.15**	**<0.001**	**-0.12**	**<0.001**	0.005	0.35
Cluster Stability	**0.98**		**0.64**		**0.57**		**0.82**		**0.83**	

Table E4.10: *Univariate logistic regression analysis using sensitization as continuous variable (IgE titer). *Coeff: The coefficient translates into a value of how likely a child is assigned to that cluster based on the variable response.*

Variable	**Cluster 1 (n=132)**		**Cluster 2 (n=339)**		**Cluster 3 (n=109)**		**Cluster 4 (n=18)**		**Cluster 5 (n=14)**	
	Coeff	p-value	Coeff	p-value	Coeff	p-value	Coeff	p-value	Coeff	p-value
Age of Onset	**-0.03**	**0.03**	**-0.21**	**<0.001**	**0.25**	**<0.001**	-0.00005	0.99	-0.004	0.48
Asthma attacks	0.005	0.74	**-0.10**	**<0.001**	**-0.03**	**0.02**	-0.003	0.52	**0.13**	**<0.001**
Total IgE	-0.003	0.86	**-0.06**	**0.003**	-0.05	0.01	**0.11**	**<0.001**	0.0005	0.99
Asthma Severity	**0.39**	**<0.001**	**-0.29**	**<0.001**	**-0.09**	**<0.001**	-0.01	0.05	-0.01	0.86
Cluster Stability	**0.97**		**0.81**		**0.67**		**0.70**		**0.67**	

4.8 Supplementary References

E1. Sekerel BE, Saraclar Y, Kalayci O, Cetinkaya F, Tuncer A, Adalioglu G. Comparison of four different measures of bronchial responsiveness in asthmatic children. Allergy 1997;52:1106-9.

E2. Joos GF, O'Connor B, Anderson SD, et al. Indirect airway challenges. Eur Respir J 2003;21:1050-68.

E3. C H. Cluster-wise assessment of cluster stability. London UK: University College London; 2006.

Chapter 5 Longitudinal trajectories of wheeze exacerbations from infancy to school age and their association with early-life risk factors and late asthma outcomes

Matea Deliu MD, Sara Fontanella PhD, Sadia Haider PhD, Matthew Sperrin PhD, Nophar Geifman PhD, Clare Murray MD, Angela Simpson MD PhD, Adnan Custovic MD PhD FAAAI

Clin Exp Allergy (under review) 2019

5.1 Rationale for this study

Using the data from a well-defined and almost complete dataset from Turkey, we identified exacerbations to be one of the key features of understanding the heterogeneity of asthma. However, it is not possible to perform any longitudinal analysis as there is no follow-up data. The Manchester Asthma and Allergy Study (MAAS) provided an ideal setting in order to utilise complex longitudinal methodology as it is a birth cohort that contains multiple follow up points, detailed data on exposures and biomarkers, genetics (though not used in this thesis), and associated detailed lung function and sensitization data. This would allow us to ascertain the trajectories of exacerbations from early onset to late childhood.

5.2 Abstract

Introduction: Exacerbation-prone asthma subtype has been reported in studies using data-driven methodologies. However, longitudinal patterns of exacerbations throughout childhood have not been studied.

Objective: To investigate distinct longitudinal trajectories of wheeze exacerbations from infancy to school-age.

Methods: We applied longitudinal k-means clustering to derive exacerbation trajectories among participants in a population-based birth cohort with confirmed exacerbations in primary healthcare records. We examined the association of derived clusters with lung function, airway hyperreactivity and inflammation, allergic sensitisation, and use of asthma medication.

Results: 498/887 children (56%) had physician-confirmed wheeze in medical records up to age 8 years, of whom 160 had at least one confirmed severe exacerbation. A two-cluster model provided the optimal solution for exacerbation trajectories among these 160 children. We assigned clusters as “Early-onset frequent exacerbations (FE)” (n=10, 6.3%) and “Infrequent exacerbations (IE)” (n=150, 93.7%). Shorter duration of breastfeeding was the strongest risk factor for FE (median weeks, 0 [IQR: 0-1.75] vs IE 6 (IQR: 0-20), p<0.001). When we compared children in the exacerbation clusters with those who never wheeze (n=389), or wheeze but have no exacerbations (n=338), the proportion of allergic sensitization was highest in those that have exacerbations. By adolescence, children who have exacerbations were more likely to have a diagnosis of asthma, require more inhaled corticosteroids, and have diminished lung function.

Conclusion: We have identified two distinct patterns of asthma exacerbations during childhood with different late-childhood asthma outcomes, early-life risk factors, and lung function. These results indicate that exacerbations represent a distinct susceptibility phenotype.

5.3 Introduction

Asthma is the most common chronic disease in children. Despite advances in treatment and changes in guidelines, severe exacerbations continue to occur, thus creating a major burden on not only the child and the child’s family, but also on health care resources.[1] Notably, a small proportion of children with frequent exacerbations account for the majority of the total exacerbation burden in childhood.[2] Recent studies utilizing data-driven methods have shown that exacerbations may provide meaningful insight into understanding asthma heterogeneity.[3,4] Identified risk factors for severe exacerbations include younger age, male sex, race, parental smoking history, socioeconomic status, diminished lung function, severe exacerbation in previous year, respiratory virus infections, and synergistic effects of allergen sensitisation, exposure and virus infection.[5-11]

Exacerbations encompass the crux of the definition of severe asthma, in that a child's asthma is considered to be severe if they've had frequent severe (use of oral steroids) or serious (hospital admission) exacerbations along with poor symptom control and airflow limitation.[12] However, although exacerbations have long been known to be an element of the severe asthma 'phenotype', it is increasingly clear that children can have asthma attacks despite good adherence with treatment and relatively good asthma control.[13,14] Studies using data-driven methods have shown that exacerbations are present within different asthma clusters of varying disease severity,[4] and that patients with mild disease can have high rates of severe exacerbations, whereby decreasing symptoms may not always mean a decrease in the number of exacerbations.[7,15] Using data driven techniques, we have recently identified an exacerbation-prone asthma cluster that contained a proportion of children with mild asthma and normal lung function.[16] This data suggests that patients with exacerbations may account for a separate susceptibility phenotype, relatively independent of whether or not a child is deemed to have severe asthma, with a likely unique aetiology.

Despite the current advancement of knowledge and treatment options in this domain, clinicians are still able to only partly prevent an impending exacerbation.[17] A recent clinical trial in children tested the hypothesis that substantially increasing the dose of inhaled corticosteroids (ICS) in the period just before an exacerbation when symptom control weakens would prevent a full-blown exacerbation.[18] However, the results were disappointing and the side effects outweighed any benefit.

We hypothesised that there are distinct patterns of severe exacerbations of wheezing during childhood, and that uncovering such patterns may help ascertain whether there are different mechanisms contributing to exacerbations. To address our hypothesis, within an unselected birth cohort, we identified longitudinal trajectories of severe exacerbations of wheezing by analysing exacerbation patterns from birth to school-age. We then identified their early life risk factors, and asthma–related outcomes and lung function measures in late childhood.

5.4 Methods

5.4.1 Study Population

The Manchester Asthma and Allergy Study is a population-based birth cohort, and is described in detail elsewhere.[2,19] Subjects were recruited prenatally and followed prospectively. The study was approved by the Local Ethics Committee. Parents provided written informed consent. For more details about methods and variable definitions please see Online supplement.

5.4.2 Data sources and definition of variables to identify exacerbation trajectories

We extracted data from electronic and paper-based primary care medical records, including prescriptions of any medication (type, dose, and indication) including oral corticosteroid prescriptions, episodes of wheeze, emergency department admissions, and asthma or wheeze-related admissions to hospital. Age in days at the time of each event was recorded.[20] This data was available from birth to age 8 years.

Severe exacerbation of wheeze/asthma: Defined from primary care records as either receipt of oral corticosteroids (OCS) for at least 3 days or hospital admission or emergency department visit because of asthma/wheeze requiring OCS.[21] We recorded age (in days) of each exacerbation to provide an accurate account of each episode.

5.4.3 Data sources and definition of variables to validate exacerbation trajectories

Children attended clinical follow-ups at ages 1, 3, 5, 8, 11, and 16 years. Validated questionnaires were interviewer-administered to collect information on symptoms, environmental exposure and treatments received. Early life risk factors, including environmental tobacco smoke (ETS) exposure, pet ownership, length of breastfeeding, day-care attendance, presence of siblings, and family history of asthma were ascertained during the last trimester of pregnancy and in the first year of life. Allergic sensitization was ascertained using skin prick tests (SPT; ages 3-16) to common inhalant and food allergens.

Current asthma at age 16 years: Defined as the presence of any two of the following three features: 1) Current wheeze; 2) Current use of asthma medication; and 3) Physician-diagnosed asthma ever.[22]

Asthma severity: We created a composite variable to represent symptom control and step in medication as based on the British Thoracic Society (BTS) guidelines.[23] BTS step 1 and 2 are considered to be 'Mild asthma', step 3 and 4 are considered as 'Moderately severe asthma', and step 5 is considered to be 'Severe asthma'.

Lung function, airway hyperreactivity and airway inflammation

We measured lung function using spirometry at ages 8, 11 and 16 years using a Lilly pneumotachograph system with animated incentive software (Jaeger, Würzburg, Germany), or for home visits, a flow turbine spirometer (Micro Medical, UK).[24] FEV_1 % predicted[25] and FEV_1/FVC ratio were recorded. Specific airway resistance (sRaw) was measured using whole-body plethysmography (Masterscreen Body 4.34; Jaeger, Würzburg, Germany) plethysmography at ages 3, 5, 8, 11, 16.[19,26] Airway hyperreactivity (AHR) was measured using standard quadrupling doses of methacholine in a 5-stage process at ages 8 and 11 years.[27] Children were considered to have AHR if there was a 20% decrease in FEV_1 by the final stage (16mg/mL). We also calculated a dose-response slope.[28] Airway inflammation was recorded at age 8, 11, and 16years as a measure of Fractional Exhaled nitric oxide (FeNO) and performed according to the American Thoracic Society guidelines using either a chemiluminescence analyser or an electrochemical analyser (NIOX, Solna, Sweden) .[29] Data were expressed in parts per billion (ppb).

5.4.4 Statistical Analysis

For ascertaining exacerbation patterns, we used primary care records data (ages 1-8) as this provided an objective account of hospital admissions, receipt of oral corticosteroids and presence of wheeze.

Exploratory analysis to identify patterns of exacerbations: We used cluster analysis for longitudinal data to identify whether there are subgroups of patients with similar exacerbation patterns. We applied a longitudinal extension of the k-means algorithm (KmL)[30] to "number of exacerbations". The KmL[30] technique is a partitional clustering method. The optimal number of clusters was assessed using the Calinski-Harabatz criterion[31]. Results for *post-hoc* longitudinal cluster analysis were obtained through the KmL package developed in R software[32]. Technical details are provided in the online supplement.

To assess differences between exacerbation clusters, we used either a t-test, χ^2 test, Mann-Whitney test, or one-way ANOVA as appropriate. To check which early-life risk factors predict trajectory membership, we used multinomial logistic regression models up to age 8.

In order to identify differences between children with different exacerbation trajectories, those with wheezing but no exacerbations, and children who have never wheezed, we used a multinomial logistic regression model with lung function, AHR, atopy, asthma severity, and asthma medication as outcomes. All analysis was performed using R software (www.r-project.org/).[33]

5.5 Results

5.5.1 Population characteristics and participant flow

Of 1184 children born into the cohort, 984 families gave consent for the review of medical records, of whom we extracted data for 887 children[20]. Of those, 498 (56%) had physician-confirmed wheeze in their medical records on at least one occasion up to age 8 years, and 389 never wheezed. Of 498 children with confirmed wheezing, 160 (32%) had at least one confirmed severe exacerbation in the first 8 years, and 338 had no exacerbations. The median age of the first recorded wheeze episode was 676 days (IQR: 187-863), and of the first exacerbation 893 days (IQR: 343-1238). The annual incidence of exacerbations ranged from 5% during ages 1-5, to 1.5% by age 8. Descriptive characteristics of 160 children who had at least one severe exacerbation of wheezing are shown in Table E5.1; 68 (43%) had a physician diagnosis of asthma in the same period, and 116 (73%) were prescribed ICS at some point. Current use of ICS among current exacerbators in each 12-month period from birth to age 8 increased from 31% in the first year of life to 79% in year 8. Among children with exacerbation(s), on average 11% per annum had ≥3 exacerbations in the preceding 12 months (frequent exacerbators); all but one of these children received asthma medication.

5.5.2 Identification of exacerbation trajectories and their associates

To identify exacerbation trajectories, we analysed data from 160 children who had at least one confirmed exacerbation from birth to age 8 years. According to the Calinski-Harabatz index (Figure E5.1), the optimal model that best described the data was a 2-class solution. Figure 5.1 shows the exacerbations patterns of the children in these two trajectories. We assigned trajectories as “Early-onset frequent exacerbations (FE)” (n=10, 6.3%, median =4) and “Infrequent exacerbations (IE)” (n=150, 93.7%, median =1).

The associations of exacerbation trajectories with risk factors and clinical features in the first 8 years of life are presented in Table E5.2. Shorter duration of breastfeeding was the strongest risk factor for FE (median weeks, 0 [IQR: 0-1.75] *vs* IE 6 (IQR: 0-20), p<0.001). Family history of asthma and tobacco smoke exposure did not differ between clusters.

Children in FE cluster were significantly more likely to have eczema in the first 3 years of life, but not thereafter. Co-morbid rhinitis and allergic sensitization did not differ between the clusters. Children in FE cluster were more likely to have persistent wheeze (90% vs 47%, p=0.03, Table E5.3), and by age 8 years, 90% of children in the FE cluster (vs. 39% in IE) had doctor diagnosed asthma (p=0.002). Although the majority of children in both clusters had mild asthma (BTS step 1 or 2), children in FE cluster accounted for the highest percentage of more severe asthma (Table E5.4). It is of note that 41% of children in IE cluster and 10% of children in FE cluster at age 3 received no asthma treatment. These figures were similar at age 5 (37% IE and 10% FE) and age 8 (25% IE, 10% FE).

We observed significant differences in lung function and airway inflammation by age 8 between the two exacerbation trajectories (Table 5.1). Children with FE had significantly lower FEV_1 % predicted and FEV_1/FVC, and significantly higher FeNO and AHR compared to IE. Similarly, sRaw was markedly and significantly higher in FE compared to IE cluster throughout childhood (Figure E5.1).

Figure 5.1: Longitudinal trajectories of exacerbations. Cluster 1: infrequent exacerbations, N=150 (93.7%). Cluster 2: Early-onset frequent exacerbations, N=10 (6.3%). Each line represents an individual trajectory.

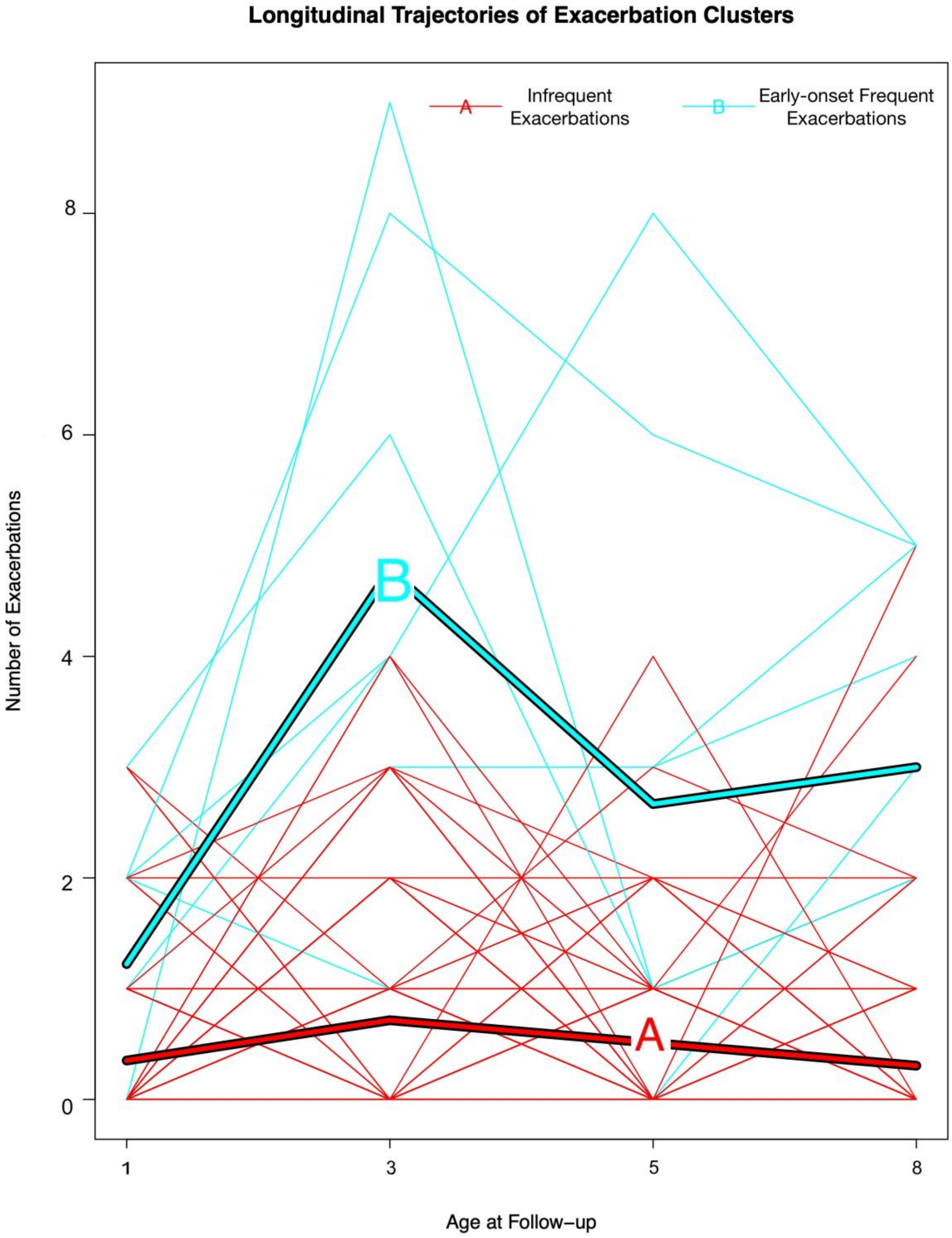


Table 5.1: *Differences in lung function, airway inflammation and airway hyper-reactivity between early onset frequent exacerbations and infrequent exacerbations. FEV$_1$= forced expiratory volume in 1 second, FeNO= fraction of exhaled nitrogen oxide; t-test used for differences between means*

	Cluster 1 (Infrequent Exacerbations); n=150	**Cluster 2 (Frequent Exacerbations); n=10**	**p-value**
FEV$_1$, mean (95%CI), age 8	95.6 (93.3-97.9)	91.1 (80.9-101.3)	**<0.001**
FEV$_1$/FVC, mean (95%CI), age 8	85.1 (83.9-86.2)	78.1 (72.8-83.4)	**<0.001**
sRAW, mean (95%CI), age 3	1.1 (0.9-1.2)	1.5 (1.3-1.6)	**<0.001**
sRAW, mean (95%CI), age 5	1.2 (1.0-1.3)	1.3 (1.1-1.4)	**<0.001**
sRAW, mean (95%CI), age 8	1.2 (1.0-1.3)	1.8 (1.5-1.9)	**<0.001**
Methacholine DRR slope, mean (95%CI), age 8	11.9 (4.9-17.4)	14.9 (2.5-21.3)	0.08
FeNO, mean (95%CI), age 8	11.5 (7.8-19.5)	58.5 (24.2-79.3)	**<0.001**

5.5.3 Comparison of children who never wheezed, wheezers with no exacerbations, and exacerbation clusters

In a further analysis, we compared children who wheezed, but have not had exacerbations (WNE), with those in the two exacerbation clusters (IE and FE), using children who never wheezed (NW) as reference (Table 5.2). The duration of breastfeeding was significantly shorter among the two exacerbation clusters, with a median of 0 weeks (RR 0.92, (95%CI: 0.85-0.99), p<0.001) in FE class. Maternal smoking during pregnancy significantly increased the risk of all three groups characterised by the presence of wheezing, with the magnitude of risk being highest among children in the FE cluster.

Figure 5.2 shows trajectories of allergic sensitization from age 1 to age 16 years across the four groups. The proportion of children who were sensitized in the FE cluster steadily increased at each age and remained markedly higher than children with infrequent exacerbations and those that have never had an exacerbation.

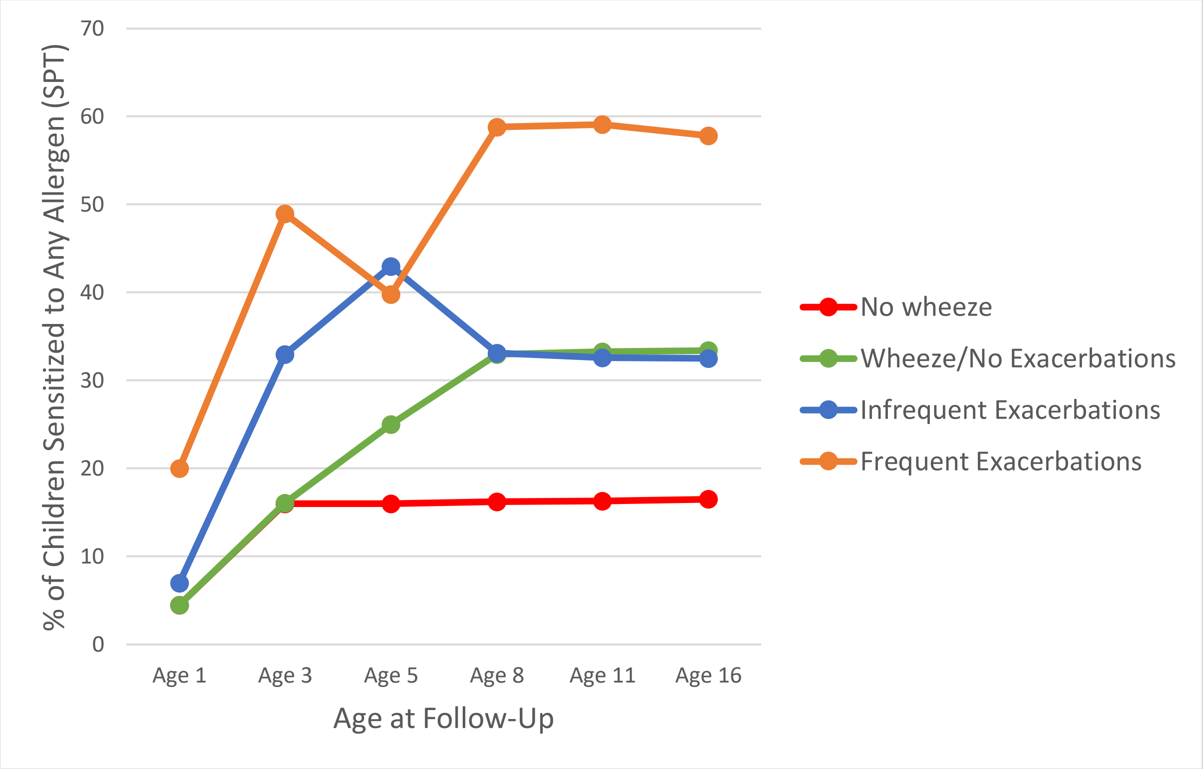
Table 5.2: *Associations of exacerbation clusters with early-life risk factors, skin test responses and co-morbidities: multinomial logistic regression using children who never wheezed (NW) as the reference. SPT: skin prick test; quantitative continuous variable presented as median (IQR); ordinal variables represented as frequencies (%); RR=relative risk; CI=confidence interval. Bold values represent significant p-values.*

	NW (reference) (n=389)	**WNE (n=338)** **RR (95%CI)** **p-value**	**IE (n=150)** **RR (95%CI)** **p-value**	**FE (n=10)** **RR (95%CI)** **p-value**
Gender (boys)	162, (42%)	166, (49%) 1.5 (1.1-2.0) 0.01	**82, (55%)** **2.5 (1.7-4.0)** **<0.001**	5, (50%) 1.2 (0.3-4.1) 0.81
Family history of asthma	91, (23%)	**121, (36%)** **1.8 (1.3-2.4)** **<0.001**	**51, (34%)** **1.6 (1.1-2.2)** **0.01**	3, (30%) 1.2 (0.4-4.3) 0.74
Younger sibling	165, (42%)	136, (40%) 0.91 (0.64-1.22) 0.54	**48, (32%)** **0.65 (0.43-0.98)** **0.04**	3, (30%) 0.76 (0.18-3.23) 0.71
Older Sibling	208, (53%)	170, (50%) 0.88(0.65-1.2) 0.39	**93, (62%)** **1.5 (1.0-2.3)** **0.03**	5, (50%) 0.87 (0.25-3.1) 0.83
Breastfeeding (weeks), median (IQR)*	10 (0-28)	8 (0-24) 0.99 (0.98-1.0) 0.08	**6 (0-20)** **0.98 (0.97-0.99)** **0.006**	**0 (0-1.8)** **0.92 (0.85-0.99)** **0.009**

Day care attendance	280, (72%)	212, (63%) 0.76 (0.55-1.1) 0.11	**78, (52%)** **0.59 (0.39-0.91)** **0.01**	7, (70%) 1.3 (0.25-6.2) 0.77
Maternal smoking during pregnancy	101, (26%)	**125, (37%)** **1.6 (1.2-1.9)** **<0.001**	**53, (35%)** **1.4 (1.1-1.9)** **0.02**	**5, (50%)** **2.8 (1.3-6.3)** **0.01**
Tobacco exposure, age 1y	96, (25%)	**116, (34%)** **1.6 (1.2-2.2)** **0.004**	46, (31%) 1.4 (0.9-2.1) 0.14	4, (40%) 2.0 (0.6-7.4) 0.28
Tobacco exposure, age 3y	89, (23%)	**116, (34%)** **1.9 (1.4-2.7)** **<0.001**	**46, (31%)** **1.8 (1.2-2.8)** **0.005**	4, (40%) 2.6 (0.7-9.9) 0.15
Tobacco exposure age 5y	91, (23%)	**119, (35%)** **1.8 (1.3-2.5)** **<0.001**	42, (28%) 1.3 (0.9-2.5) 0.19	5, (50%) 3.2 (0.9-11.3) 0.07
Atopic sensitization (SPT), age 3y	65, (17%)	61, (18%) 1.2 (0.8-1.8) 0.4	**50, (33%)** **3.2 (2.1-5.1)** **<0.001**	**5, (50%)** **10.9 (2.1-57.7)** **0.004**
Atopic sensitization (SPT), age 5y	75, (19%)	95, (28%) 1.6 (1.2-2.3) 0.005	**67, (45%)** **3.4 (2.2-5.1)** **<0.001**	**4, (40%)** **4.8 (1.0-22.2)** **0.04**

Dog ownership, birth	54, (14%)	62, (18%) 1.4 (0.9-2.1) 0.09	20, (13%) 1.0 (0.6-1.8) 0.88	3, (30%) 3.1 (0.7-12.8) 0.12
Cat ownership, birth	72, (18%)	76 , (22%) 1.3 (0.9-1.9) 0.15	24, (16%) 0.9 (0.5-1.5) 0.70	4, (40%) 3.5 (0.9-13.4) 0.06
Rhinitis, age 5y	65, (17%)	**103, (30%)** **2.3 (1.6-3.2)** **<0.001**	**55, (37%)** **3.2 (2.0-4.9)** **<0.001**	**5, (50%)** **7.9 (1.8-34.2)** **0.005**
Eczema, age 1y	119, (31%)	122, (36%) 1.3 (0.9-1.8) 0.06	**54, (36%)** **1.6 (1.1-2.5)** **0.02**	**7, (70%)** **15.3 (1.9-126.2)** **0.01**
Eczema, age 3y	78, (20%)	81, (24%) 1.3 (0.9-1.9) 0.13	**47, (31%)** **2.3 (1.5-3.5)** **<0.001**	**6, (60%)** **11.5 (2.3-58.1)** **0.004**
Eczema, age 5y	105, (27%)	105, (31%) 1.2 (0.9-1.7) 0.26	**56, (37%)** **1.6 (1.1-2.5)** **0.01**	4, (40%) 2.6 (0.6-10.4) 0.19

Figure 5.2: *Trajectories of atopic sensitization from age 1- age 16 years among children who never wheezed, those who wheeze but had no exacerbations, and two exacerbation clusters*



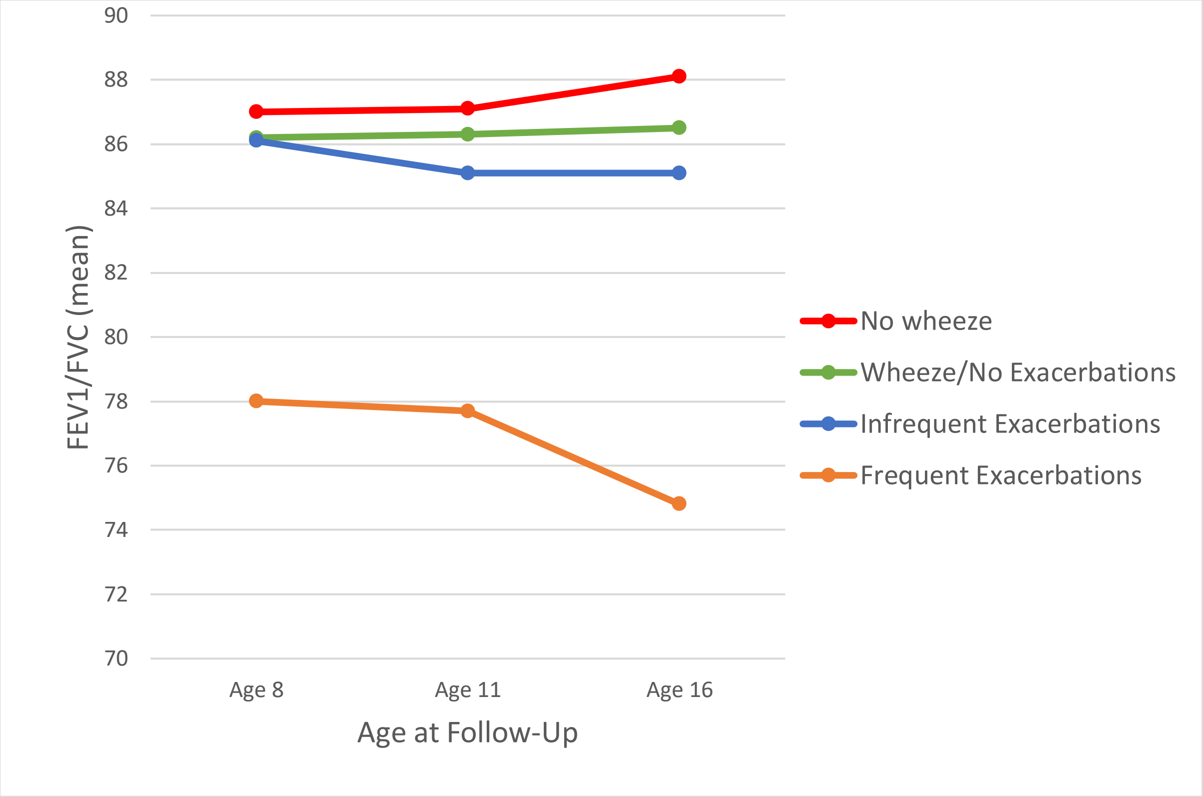
5.5.4 Late lung function and asthma outcomes

Trajectories of lung function from age 8 to age 16 years in the four groups are shown in Figure 5.3. Compared to NW, children with wheeze and/or exacerbations had significantly diminished lung function from mid-school age to adolescence (Figure 5.3, Table E5.5). Table 5.3 shows lung function and FeNO at age 16 years in the four groups and Figure E5.3 shows trajectories of airway inflammation from age 8 to age 16 years. FEV_1/FVC was significantly lower in the FE cluster compared to all other groups, while FeNO was significantly increased in both exacerbation groups.

At age 16 years, children who had exacerbations were more likely to have a diagnosis of asthma than those who wheezed alone (80% in FE and 52% IE vs 25% WNE, p<0.001) (Table 5.4). Based on BTS guideline treatment steps, children with exacerbations were more likely to have moderately-severe asthma, particularly those in the FE cluster. Similarly, the proportion of children on inhaled corticosteroids was significantly higher in children with exacerbations than those who wheezed but did not exacerbate.

Figure 5.3: *Trajectories of lung function from age 8 to age 16 years among children who never wheezed, those who wheeze but had no exacerbations, and two exacerbation clusters*

a) FEV$_1$/FVC % predicted



b) FEV$_1$% predicted

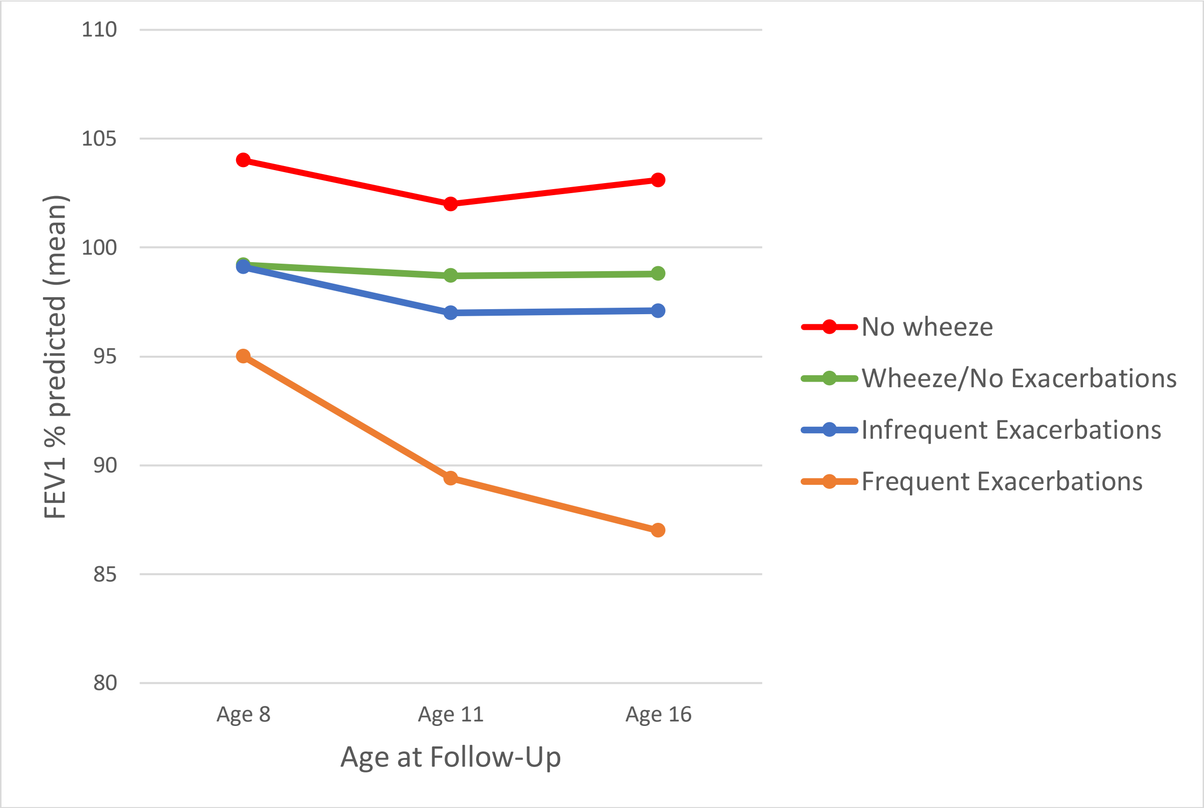

110
105
100
95
90
85
80
FEV1 % predicted (mean)
Age 8
Age 11
Age 16
Age at Follow-Up
No wheeze
Wheeze/No Exacerbations
Infrequent Exacerbations
Frequent Exacerbations

***Table 5.3**: Lung function at age 16 years among children who never wheezed (NW), those who wheezed, but have not had exacerbations (WNE), and children in the two exacerbation clusters (IE and FE). Children with lung function tests, N=559. See online supplement for visualisation. FEV$_1$= forced expiratory volume in 1 second, FeNO= fraction of exhaled nitrogen oxide. Quantitative variables represented as mean (95% confidence interval).*

	NW (reference) **(n=247)**	**WNE (n=217)** **N (%)**	**IE (n=88)** **N (%)**	**FE (n=7)** **N (%)**	**p-value**
FEV$_1$, mean (95%CI), age 16y	101.4 (99.8-102.1)	97.9 (96.5-99.2)	95.5 (93.1-97.9)	87.3 (78.9-95.7)	0.64
FEV$_1$/FVC, mean (95%CI), age 16y	89.9 (89.3-90.5)	88.1 (87.3-88.8)	85.1 (83.4-86.7)	74.7 (61.5-87.8)	**0.001**
sRAW, mean (95%CI), age 16y	0.9 (0.91-1.0)	1.0 (0.9-1.01)	1.0 (0.9-1.1)	1.5 (1.4-1.6)	0.09
FeNO, mean (95%CI), age 16y	21.7 (19.7-23.8)	30.9 (27.4-34.4)	43.1 (34.6-51.5)	52.6 (19.9-85.3)	**0.01**

Table 5.4: *Asthma-related outcomes at age 16 years among children who wheezed, but have not had exacerbations (WNE), and children in the two exacerbation clusters (IE and FE). *Fisher's exact test for asthma treatment due to small numbers. Chi-squared used for binary data.*

	WNE (n=244) **N (%)**	**IE (n=97)** **N (%)**	**FE (n=9)** **N (%)**	**p-value**
Current asthma, age 16 years	32, (13%)	29, (30%)	6, (67%)	**0.02**
Use of inhaled corticosteroids at age 16 years	45, (18%)	37, (15%)	7, (77%)	**0.04**
Ever doctor-diagnosed asthma by age 16y	86, (35%)	78, (89%)	8, (89%)	**<0.001**
Asthma severity				
No asthma treatment	202, (83%)	61, (69%)	1, (11%)	**<0.001**
Step 1	24, (9%)	15, (15%)	1, (11%)	
Step 2	14, (6%)	14, (14%)	3, (33%)	
Step 3	4, (2%)	7, (7%)	4, (44%)	

5.6 Discussion

5.6.1 Key findings

To the best of our knowledge, this is the first study that has mapped trajectories of numbers of exacerbations over time. In our birth cohort study, with the use of data-driven methodology, we identified two distinct trajectories of exacerbations of wheezing in childhood with different outcomes, early-life risk factors, and lung function. We have shown that the frequent exacerbations subtype has an early high-intensity onset which then subsides somewhat, but then persists through childhood. Infrequent exacerbation subtype tends to be stable throughout. Both these subtypes are associated with the persistent wheeze phenotype. In a further analysis comparing children with exacerbations and children who wheeze but do not exacerbate, children with exacerbations were more likely to be sensitized to any allergen, have a diagnosis of asthma, and require more inhaled corticosteroids for symptom control. These children also had significantly diminished lung function and greater airway inflammation by late childhood. Despite this, we also found that a large subset of children that exacerbate also have mild asthma with good symptom control, normal lung function and relatively low medication use.

5.6.2 Limitations and strengths

One limitation of our study arises from the number of children included in the initial analysis to determine the longitudinal trajectories. Given the relatively small number of children in FE class, the study may not have picked up all significant differences, but rather only the highly significantly ones. However, exacerbations are rare events and children who exacerbate make up 18% of our population, which is similar to other cohorts.[1,34] We did not have a replication population, as another birth cohort with transcribed data from health-care records could not be identified to allow objective ascertainment of severe exacerbations. Future research would include validating our results in other populations.

A major strength of our study is that the data was extracted from primary care records which presents the most accurate account of exacerbation frequency. Our primary care records also contain precise information on prescriptions. We have not had to rely on the recollection of parents. Another strength is the data-driven methodological approach used in a large well-defined prospective birth cohort study which has removed any a priori assumptions.

5.6.3 Interpretation

This study builds upon our previously published work in a well-defined patient cohort.[3] Using similar methodology, we identified an exacerbation-prone asthma subtype, independent of asthma severity or control, as one of the key features for disaggregating asthma and identifying disease subtypes. Although there is general consensus that there are different asthma subtypes, only focusing on key determinants of asthma presence are not enough to uncover the heterogeneity.

Duration of breastfeeding was the most significant early life risk factor for exacerbations, particularly frequent exacerbations. Children from mothers who had the shortest duration of breastfeeding were more likely to experience exacerbations in childhood. Various studies have reported the protective nature of breastfeeding on asthma development, or reduction in severity[35-37], including a recent study by Ahmadizar *et al* showing a lower risk of exacerbations later in life.[38] These results suggest that breastfeeding may have an immunomodulatory effect on exacerbations. Breastmilk contains high levels of immunoglobulins, lactoferrin, cytokines, and prebiotic structures that influence the development of immune systems.[39] Furthermore, it has been shown that breastmilk can alter the composition of the gut microbiome.[40] However, further prospective research is required to clarify the underlying mechanisms. Other early-life risk factors such as day care attendance, position in sibship, and exposure to tobacco smoke were significantly associated with exacerbations throughout childhood. These children were also more likely to have co-morbid rhinitis and eczema as well as being allergically sensitized. In particular, children who exacerbate were significantly and markedly more likely to be sensitized to any allergen throughout childhood and adolescent years. This is in concordance to a similar allergy-driven exacerbation cluster identified by the Trousseau Asthma Programme which could suggest that similar risk factors may be contributing to atopy and exacerbations[41]

Looking more closely at the children who exacerbate in childhood within our cohort, it is evident that they are more likely to have the persistent wheeze phenotype. This could be explained in two ways: either our children were not on adequate therapy (only 50% of children in with FE were on inhaled corticosteroids by age 8) and were therefore having exacerbations, or, that these children were on correct therapy at the time of follow-up, but still exacerbated due to some other underlying mechanism. The fact that many children with well-controlled asthma and normal lung function still frequently exacerbate emphasizes the need for directed research into the pathophysiological mechanisms underlying this process as we are still unable to prevent impending exacerbations.[42] Furthermore, increasing inhaled corticosteroid therapy in order to prevent exacerbations or prolong the time to first exacerbation has been shown to be ineffective.[18]

The proportion of children with the most 'severe' asthma in our cohort was highest in children who had frequent exacerbations. Our results also show that a large subset of children that exacerbate have mild asthma with good symptom control and relatively low medication use. This provides evidence that exacerbation-prone asthma may represent a separate subtype, not necessarily associated with severe disease or severe airway obstruction. Other recent studies using similar approaches have added value to the notion of a separate exacerbation-prone subtype independent of asthma severity or control.[3,7]

We found that exacerbations were associated with more airway inflammation, airway hyper-responsiveness, and poorer lung function, which is in line with previous studies.[4,7,34] These effects were even more pronounced in those children that had frequent exacerbations. It is interesting to note that the majority of children with exacerbations had normal FEV_1. However, when we looked at the ratio of FEV_1/FVC, which is a measure of airway obstruction, these values dropped significantly in children who had frequent exacerbations, indicating that FEV_1/FVC ratio may be a better predictor of those at risk of exacerbations in this age group. Nevertheless, it is evident that children who exacerbate have poorer lung function when compared to those who just wheeze. In a recent longitudinal multi-cohort study, children with recurrent early life exacerbations were more likely to persist in having poor lung function by adulthood[43] which is an important factor in the development of chronic obstructive pulmonary disease (COPD).[44,45] Therefore targeting the prevention of exacerbations may alter this persistently low lung function in order to prevent ongoing decline by the time the physiological plateau is reached in early adulthood.

Studies have shown that viral respiratory infections, particularly rhinoviruses, are major triggers of asthma exacerbations by mounting a large type-2 inflammatory response mediated by eosinophils.[8] A recent study which used machine learning applied to 28 rhinovirus-16 induced cytokines and chemokines induced by stimulation of blood mononuclear cells of children described six immunophenotypes of anti-virus responses[46]. The IFN^{lowest}Inflam^{high}Th2-chem^{low}Reg^{mod} cluster (lowest interferon induction and highest proinflammatory cytokine response) was associated with early-onset asthma and sensitization, and the highest risk of asthma exacerbations[46]. These characteristics resemble other exacerbation prone asthma subtypes.[34,41] However, treatments to reduce eosinophils, and thereby the surmounting inflammatory response, using corticosteroids and biologics have not been able to fully prevent exacerbations, suggesting that other cells and mechanisms are likely involved. New data from Touissant *et al* has provided some new insight into the role of neutrophils and the formation of neutrophil extracellular traps as major contributors in initiating viral induced asthma exacerbations, perhaps suggesting a novel therapeutic target.[47]

In conclusion, we have identified two distinct patterns of asthma exacerbations during childhood with different late-childhood asthma outcomes, early-life risk factors, and lung function when compared to children who never wheeze and those that wheeze, but have no exacerbations. These results indicate that exacerbations represent an independent susceptibility phenotype. Furthermore, ascertaining the stability of exacerbation frequency in asthma is important, as it could help identify patients at risk of having future exacerbations. This would encourage more in-depth initial clinical evaluations, closer longitudinal follow-up, optimizing adherence to currently available treatment strategies, or the development of novel prevention strategies.

5.7 References

1. Suruki RY, Daugherty JB, Boudiaf N, Albers FC. The frequency of asthma exacerbations and healthcare utilization in patients with asthma from the UK and USA. *BMC Pulm Med* 2017; **17**(1): 74.
2. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *Am J Respir Crit Care Med* 2013; **188**(11): 1303-12.
3. Deliu M, Yavuz TS, Sperrin M, et al. Features of asthma which provide meaningful insights for understanding the disease heterogeneity. *Clin Exp Allergy* 2017.
4. 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**(4): 315-23.
5. Tattersfield AE, Postma DS, Barnes PJ, et al. Exacerbations of asthma: a descriptive study of 425 severe exacerbations. The FACET International Study Group. *Am J Respir Crit Care Med* 1999; **160**(2): 594-9.
6. ten Brinke A, Sterk PJ, Masclee AA, et al. Risk factors of frequent exacerbations in difficult-to-treat asthma. *Eur Respir J* 2005; **26**(5): 812-8.
7. Kupczyk M, ten Brinke A, Sterk PJ, et al. Frequent exacerbators--a distinct phenotype of severe asthma. *Clin Exp Allergy* 2014; **44**(2): 212-21.
8. Jackson DJ, Johnston SL. The role of viruses in acute exacerbations of asthma. *J Allergy Clin Immunol* 2010; **125**(6): 1178-87; quiz 88-9.
9. Haselkorn T, Zeiger RS, Chipps BE, et al. Recent asthma exacerbations predict future exacerbations in children with severe or difficult-to-treat asthma. *J Allergy Clin Immunol* 2009; **124**(5): 921-7.
10. Miller MK, Lee JH, Miller DP, Wenzel SE, Group TS. Recent asthma exacerbations: a key predictor of future exacerbations. *Respir Med* 2007; **101**(3): 481-9.
11. Murray CS, Poletti G, Kebadze T, et al. Study of modifiable risk factors for asthma exacerbations: virus infection and allergen exposure increase the risk of asthma hospital admissions in children. *Thorax* 2006; **61**(5): 376-82.
12. Chung KF, Wenzel SE, Brozek JL, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *Eur Respir J* 2014; **43**(2): 343-73.
13. Reddel H, Ware S, Marks G, Salome C, Jenkins C, Woolcock A. Differences between asthma exacerbations and poor asthma control. *Lancet* 1999; **353**(9150): 364-9.
14. Taylor DR, Bateman ED, Boulet LP, et al. A new perspective on concepts of asthma severity and control. *Eur Respir J* 2008; **32**(3): 545-54.
15. O'Byrne PM, Barnes PJ, Rodriguez-Roisin R, et al. Low dose inhaled budesonide and formoterol in mild persistent asthma: the OPTIMA randomized trial. *Am J Respir Crit Care Med* 2001; **164**(8 Pt 1): 1392-7.
16. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of Asthma Subtypes Using Clustering Methodologies. *Pulm Ther* 2016; **2**: 19-41.
17. Sorkness CA, Lemanske RF, Jr., Mauger DT, et al. Long-term comparison of 3 controller regimens for mild-moderate persistent childhood asthma: the Pediatric Asthma Controller Trial. *J Allergy Clin Immunol* 2007; **119**(1): 64-72.

18. Jackson DJ, Bacharier LB, Mauger DT, et al. Quintupling Inhaled Glucocorticoids to Prevent Childhood Asthma Exacerbations. *N Engl J Med* 2018; **378**(10): 891-901.

19. Custovic A, Simpson BM, Murray CS, et al. The National Asthma Campaign Manchester Asthma and Allergy Study. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* 2002; **13 Suppl 15**: 32-7.

20. Semic-Jusufagic A, Belgrave D, Pickles A, et al. Assessing the association of early life antibiotic prescription with asthma exacerbations, impaired antiviral immunity, and genetic variants in 17q21: a population-based birth cohort study. *Lancet Respir Med* 2014; **2**(8): 621-30.

21. Reddel HK, Taylor DR, Bateman ED, et al. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am J Respir Crit Care Med* 2009; **180**(1): 59-99.

22. Lodrup Carlsen KC, Roll S, Carlsen KH, et al. Does pet ownership in infancy lead to asthma or allergy at school age? Pooled analysis of individual participant data from 11 European birth cohorts. *PLoS One* 2012; **7**(8): e43214.

23. James DR, Lyttle MD. British guideline on the management of asthma: SIGN Clinical Guideline 141, 2014. *Arch Dis Child Educ Pract Ed* 2016; **101**(6): 319-22.

24. Beydon N, Davis SD, Lombardi E, et al. An official American Thoracic Society/European Respiratory Society statement: pulmonary function testing in preschool children. *Am J Respir Crit Care Med* 2007; **175**(12): 1304-45.

25. Stanojevic S, Wade A, Cole TJ, et al. Spirometry centile charts for young Caucasian children: the Asthma UK Collaborative Initiative. *Am J Respir Crit Care Med* 2009; **180**(6): 547-52.

26. Lowe LA, Woodcock A, Murray CS, Morris J, Simpson A, Custovic A. Lung function at age 3 years: effect of pet ownership and exposure to indoor allergens. *Arch Pediatr Adolesc Med* 2004; **158**(10): 996-1001.

27. Crapo RO, Casaburi R, Coates AL, et al. Guidelines for methacholine and exercise challenge testing-1999. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. *Am J Respir Crit Care Med* 2000; **161**(1): 309-29.

28. Beydon N, Pin I, Matran R, et al. Pulmonary function tests in preschool children with asthma. *Am J Respir Crit Care Med* 2003; **168**(6): 640-4.

29. 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**(8): 912-30.

30. Genolini C, Falissard B. KmL: a package to cluster longitudinal data. *Comput Methods Programs Biomed* 2011; **104**(3): e112-21.

31. Genolini CaF, B. Kml: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine* 2011; **104**(3): 112-21.

32. Genolini CA MS, M.; Arnaud, C. Kml and kml3d: R Packages to cluster longitudinal data. *Journal of Statistical Software* 2015; **65**(4): 1-34.

33. Team R. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.

34. Denlinger LC, Phillips BR, Ramratnam S, et al. Inflammatory and Comorbid Features of Patients with Severe Asthma and Frequent Exacerbations. *Am J Respir Crit Care Med* 2017; **195**(3): 302-13.
35. Azad MB, Vehling L, Lu Z, et al. Breastfeeding, maternal asthma and wheezing in the first year of life: a longitudinal birth cohort study. *Eur Respir J* 2017; **49**(5).
36. Dogaru CM, Nyffenegger D, Pescatore AM, Spycher BD, Kuehni CE. Breastfeeding and childhood asthma: systematic review and meta-analysis. *Am J Epidemiol* 2014; **179**(10): 1153-67.
37. Elbert NJ, van Meel ER, den Dekker HT, et al. Duration and exclusiveness of breastfeeding and risk of childhood atopic diseases. *Allergy* 2017; **72**(12): 1936-43.
38. Ahmadizar F, Vijverberg SJH, Arets HGM, et al. Breastfeeding is associated with a decreased risk of childhood asthma exacerbations later in life. *Pediatr Allergy Immunol* 2017; **28**(7): 649-54.
39. Turfkruyer M, Verhasselt V. Breast milk and its impact on maturation of the neonatal immune system. *Curr Opin Infect Dis* 2015; **28**(3): 199-206.
40. Abballe A, Ballard TJ, Dellatte E, et al. Persistent environmental contaminants in human milk: concentrations and time trends in Italy. *Chemosphere* 2008; **73**(1 Suppl): S220-7.
41. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. *Eur Respir J* 2012; **40**(1): 55-60.
42. Saglani S, Custovic A. Childhood Asthma: Advances Using Machine Learning and Mechanistic Studies. *Am J Respir Crit Care Med* 2018.
43. Belgrave DCM, Granell R, Turner SW, et al. Lung function trajectories from pre-school age to adulthood and their associations with early life factors: a retrospective analysis of three population-based birth cohort studies. *Lancet Respir Med* 2018.
44. Lange P, Celli B, Agusti A, et al. Lung-Function Trajectories Leading to Chronic Obstructive Pulmonary Disease. *N Engl J Med* 2015; **373**(2): 111-22.
45. Martinez FD. Early-Life Origins of Chronic Obstructive Pulmonary Disease. *N Engl J Med* 2016; **375**(9): 871-8.
46. Custovic A, Belgrave D, Lin L, et al. Cytokine Responses to Rhinovirus and Development of Asthma, Allergic Sensitization, and Respiratory Infections during Childhood. *Am J Respir Crit Care Med* 2018; **197**(10): 1265-74.
47. Toussaint M, Jackson DJ, Swieboda D, et al. Host DNA released by NETosis promotes rhinovirus-induced type-2 allergic asthma exacerbation. *Nat Med* 2017; **23**(6): 681-91.

5.8 Supplementary Material

5.8.1 Methods

Study design: Unselected birth cohort

Setting: A mixed urban-rural population within 50 square miles of South Manchester and Cheshire, United Kingdom located within the maternity catchment area of Wythenshawe and Stepping Hill Hospitals

Screening and recruitment: All pregnant women were screened for eligibility at antenatal visits (8-10th week of pregnancy). Of the 1499 couples who met the inclusion criteria (≤10 weeks of pregnancy, maternal age ≥18 years, and questionnaire and skin prick data test available for both parents), 288 declined to take part in the study and 27 were lost to follow-up between recruitment and the birth of a child. A total of 1184 children born into the study had at least some evaluable data.

Data from primary care medical records: Eligible GP practices were invited to participate in the study by postal information packs and telephone calls. Data access and manual extraction were performed during arranged visits to each GP practice. A trained paediatrician extracted data from electronic and paper-based primary care medical records, including prescriptions, acute wheeze episodes, hospital admissions for asthma/wheeze, oral steroid prescriptions. Timing, type of visit, symptoms, indication and prescriptions were noted for each encounter.

Definition of variables

Wheeze phenotypes[1,2]*:* (1) No wheezing: no wheeze ever; (2) Transient early wheezing: wheezing during the first 3 years, no wheezing in the previous 12 months at subsequent follow-ups; (3) Late-onset wheezing: no wheeze during the first 3 years, reported wheezing in the previous 12 months at age 5 years or later; and (4) Persistent wheezing: wheezing throughout childhood.

Current rhinitis: Positive answer to “In the past 12 months, has your child had a problem with sneezing or a runny or blocked nose when he/she did not have a cold or the flu?”

Current eczema: Positive answer to “Has your child had eczema within the past 12 months?”

Atopic sensitisation: Mean diameter (MWD) 3mm larger than the negative control to at least one allergen.

Current wheeze: physician confirmed wheeze as documented in primary care records available each year up to age 8. From ages 11-16, current wheeze is documented as a positive answer to the question “Has your child had wheezing or whistling in the chest in the last 12 months?”

Asthma medication: The use of inhaled corticosteroids and/or other asthma medication was recorded in primary care records.

Statistical Analysis

Identification of Exacerbation clusters

We used a k-means longitudinal model to ascertain the longitudinal trajectories of exacerbations in childhood. The KmL[4] technique belongs to the class of partitional clustering. The main advantages of these methods are that no distributional assumptions within clusters are required, no assumptions regarding the shape of the trajectories are made, and they are independent from time-scaling.

Formally, consider a set *S* of *n* subjects and let x_{it} be the value of variable X measured for each subject i at time t. The sequence $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{it})$ is then called a *trajectory*. The aim of KmL is, then, to divide *S* into *g* homogeneous sub-groups. Several distance measures can be chosen to assign trajectories to clusters. We used the Manhattan distance, which is more robust to outliers[4]. The optimal number of clusters is assessed using the Calinski and Harabatz criterion, which evaluates cluster validity based on the average between- and within-cluster sum of squares. Using all time points from 1-8, two optimal models were identified: a 2-cluster solution and a 4-cluster solution, with a slight preference to a 2-cluster solution (Figure E1). We then collapsed the time points to 1-3-5-8 years in order to reduce the variability and to correspond to clinical follow up.

To test the validity of exacerbation classes, we created a separate dataset that included the entire population with complete data (n=887). We allocated children to 2 extra a priori identified subgroups which we identified as “children with no wheeze at baseline” and “children with wheeze but no exacerbations”. We compared early and late childhood risk factors and lung function between the exacerbation subgroups and the aforementioned a priori ones.

5.8.2 Results

Figure E5.1: *Selection of number of clusters using the Calinski-Harabatz method.*

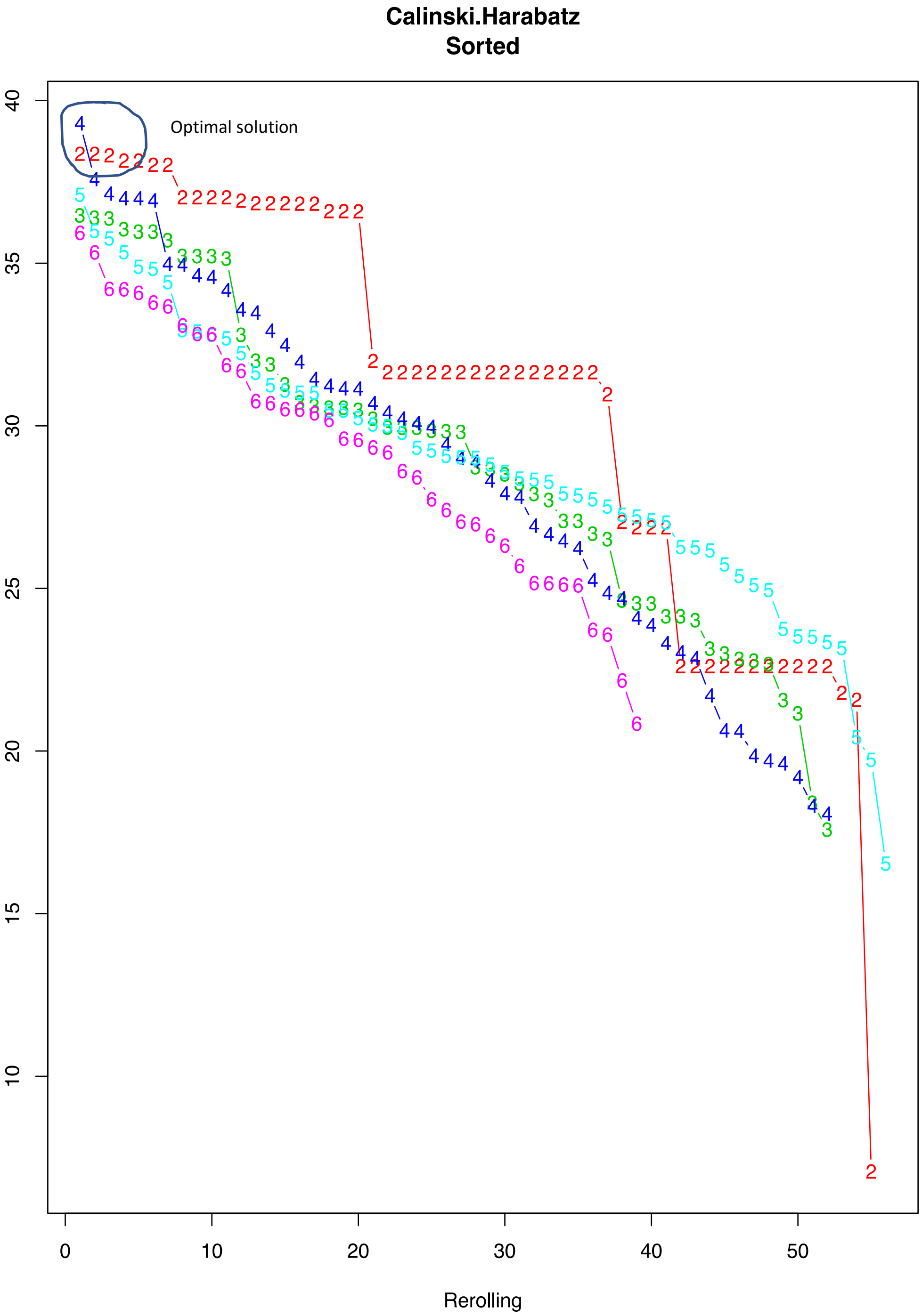


Table E5.1: *Descriptive characteristics of children who ever had an exacerbation based on primary care records. ICS: inhaled corticosteroid*

Age (years), n=160		**0-1**	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**
Current ICS in current exacerbators		14/45, 31%	18/44, 41%	20/44, 45%	27/54, 50%	20/30, 67%	14/20, 70%	10/20, 50%	11/14, 79%
Current asthma medication in current exacerbators	All exacerbations	33/45, 73%	36/44, 82%	41/44, 93%	47/54, 87%	27/30, 90%	19/20, 95%	17/20, 85%	12/14, 86%
	≥3	5/5, 100%	6/7, 86%	10, 100%	1/1, 100%	2/2, 100%	1/1, 100%	2/2, 100%	1/1, 100%
Gender (boys)		108/160, 67%							
Ever asthma in ever exacerbators		68/160, 43%							
Ever ICS in ever exacerbators		116/160, 73%							

Table E5.2: *Early-life characteristics and clinical features of exacerbation clusters. IF: Infrequent exacerbations; FE: Early-onset frequent exacerbations; SPT: skin prick test, Quantitative variable presented as median (IQR), Ordinal variables represented as frequencies (%). *Mann-Whitney test for medians. Chi-squared for binary variables*

	Cluster 1 (IF) n=150	**Cluster 2 (FE) n=10**	**p-value**
Gender (boys)	103, (67%)	5, (50%)	0.38
Family history of asthma	51, (34%)	3, (30%)	0.86
Younger sibling	48, (32%)	3, (30%)	0.84
Older sibling	93, (62%)	5, (50%)	0.59
Breastfeeding (weeks), median (IQR)*	**6 (0-20)**	**0 (0-1.75)**	**<0.001**
Day care attendance	78, (52%)	7, (70%)	0.36
Tobacco exposure, birth	53, (35%)	5, (50%)	0.64
Tobacco exposure 1y	46, (30%)	4, (40%)	0.80
Tobacco exposure 3y	46, (30%)	4, (40%)	0.88
Tobacco exposure 5y	42, (28%)	5, (50%)	0.31
Atopic sensitization (SPT), age 3y	50, (33%)	5, (50%)	0.27
Atopic sensitization (SPT), age 5y	67, (45%)	4, (40%)	0.94
Atopic sensitization (SPT), age 8y	69, (46%)	6, (60%)	0.83
Dog ownership, birth	20, (13%)	3, (30%)	0.29
Cat ownership, birth	24, (16%)	4, (40%)	0.34
Presence of rhinitis, age 5y	55, (37%)	5, (50%)	0.36
Presence of rhinitis, age 8y	51, (34%)	6, (60%)	0.13
Presence of eczema, age 1y	54, (36%)	7, (70%)	**0.03**
Presence of eczema, age 3y	47, (31%)	6, (60%)	**0.03**
Presence of eczema, age 5y	56, (37%)	4, (40%)	0.54
Presence of eczema, age 8y	48, (32%)	5, (50%)	0.24
Doctor-diagnosed asthma ever	**59, (39%)**	**9, (90%)**	**0.002**
Use of inhaled corticosteroids, age 3y	**33, (22%)**	**8, (80%)**	**<0.001**
Use of inhaled corticosteroids, age 5y	46, (31%)	6, (60%)	0.11
Use of inhaled corticosteroids age 8y	41, (27%)	5, (50%)	0.13

Table E5.3: *Wheeze phenotypes and their association with exacerbation clusters. IF: Infrequent exacerbations; FE: Early-onset frequent exacerbations; Chi-squared for binary variables*

Wheeze phenotype	**Cluster 1 (IF) n=150**	**Cluster 2 (FE) n=10**	**p-value**
No wheeze	7, (21%)	0, (0)	0.65
Late onset wheeze	50, (33%)	1, (10%)	
Transient early wheeze	15, (10%)	0, (0%)	
Persistent wheeze	70, (47%)	9, (90%)	

Table E5.4: *The distribution of asthma severity by exacerbation clusters. IF: Infrequent exacerbations; FE: Early-onset frequent exacerbations; BTS: British Thoracic Society step in asthma treatment. **Fisher's exact test used due to low numbers. Bolded values represent significant p-values*

	Cluster 1 (IF) n=150	**Cluster 2 (FE) n=10**	**p-value**
BTS age 3y**			
No asthma treatment	62, (41%)	1, (10%)	**0.04**
Step 1	48, (32%)	3, (30%)	
Step 2	31, (21%)	5, (50%)	
Step 3	1, (0.7%)	1, (10%)	
BTS age 5y**			
No asthma treatment	55, (37%)	1, (10%)	0.19
Step 1	33, (22%)	3, (30%)	
Step 2	40, (27%)	5, (50%)	
Step 3	6, (4%)	1, (10%)	
BTS age 8y**			
No asthma treatment	37, (25%)	1, (10%)	0.31
Step 1	18, (12%)	3, (30%)	
Step 2	32, (21%)	4, (40%)	
Step 3	8, (5%)	1, (10%)	

***Table E5.5:** Associations of exacerbation clusters with lung function: multinomial logistic regression using children who never wheezed (NW) as the reference. FEV_1= forced expiratory volume in 1 second, FVC= forced vital capacity, FeNO= fraction of exhaled nitrogen oxide*

	NW (reference) (n=389)	**WNE (n=338)** **RR (95%CI)** **p-value**	**IE (n=150)** **RR (95%CI)** **p-value**	**FE (n=10)** **RR (95%CI)** **p-value**
FEV_1, mean (SD), age 8y	103.9 (11.9)	**98.9 (11.7)** **0.97 (0.95-0.98)** **<0.001**	**95.6 (14.6)** **0.95 (0.03-0.97)** **<0.001**	**91.1 (14.6)** **0.93 (0.87-0.99)** **0.02**
FEV_1/FVC, mean (SD), age 8	87.9 (5.2)	**86.3 (5.8)** **0.94 (0.91-0.97)** **0.001**	**85.1 (7.6)** **0.91 (0.88-0.95)** **<0.001**	**78.1 (6.9)** **0.80 (0.71-0.89)** **<0.001**
sRAW, mean (SD), age 3y	1.1 (0.2)	**1.2 (0.2)** **2.6 (1.1-6.3)** **0.04**	**1.2 (0.3)** **12.5 (4.3-36.3)** **<0.001**	**1.5 (0.3)** **86.7 (1.1-743.2)** **<0.001**
sRAW, mean (SD), age 5y	1.1 (0.2)	**1.2 (0.2)** **5.3 (2.5-11.2)** **<0.001**	**1.2 (0.2)** **9.8 (4.1-23.7)** **<0.001**	**1.3 (0.2)** **77.8 (11.4-532)** **<0.001**
sRAW, mean (SD), age 8y	1.1 (0.2)	**1.2 (0.2)** **1.99 (1.07-3.73)** **0.03**	**1.2 (0.3)** **4.4 (2.0-9.3)** **<0.001**	**1.8 (0.2)** **62.5 (11.1-353.4)** **<0.001**
FeNO, mean (SD), age 8	9.4 (13.3)	9.6 (20.3) 1.01 (0.99-1.03) 0.08	**11.5 (19.5)** **1.02 (1.00-1.03)** **0.01**	**65.3 (31.0)** **1.05 (1.02-1.08)** **<0.001**

Figure E5.2: *Trajectories of sRAW in from age 3 to age 8 years among children in two exacerbation trajectories.*

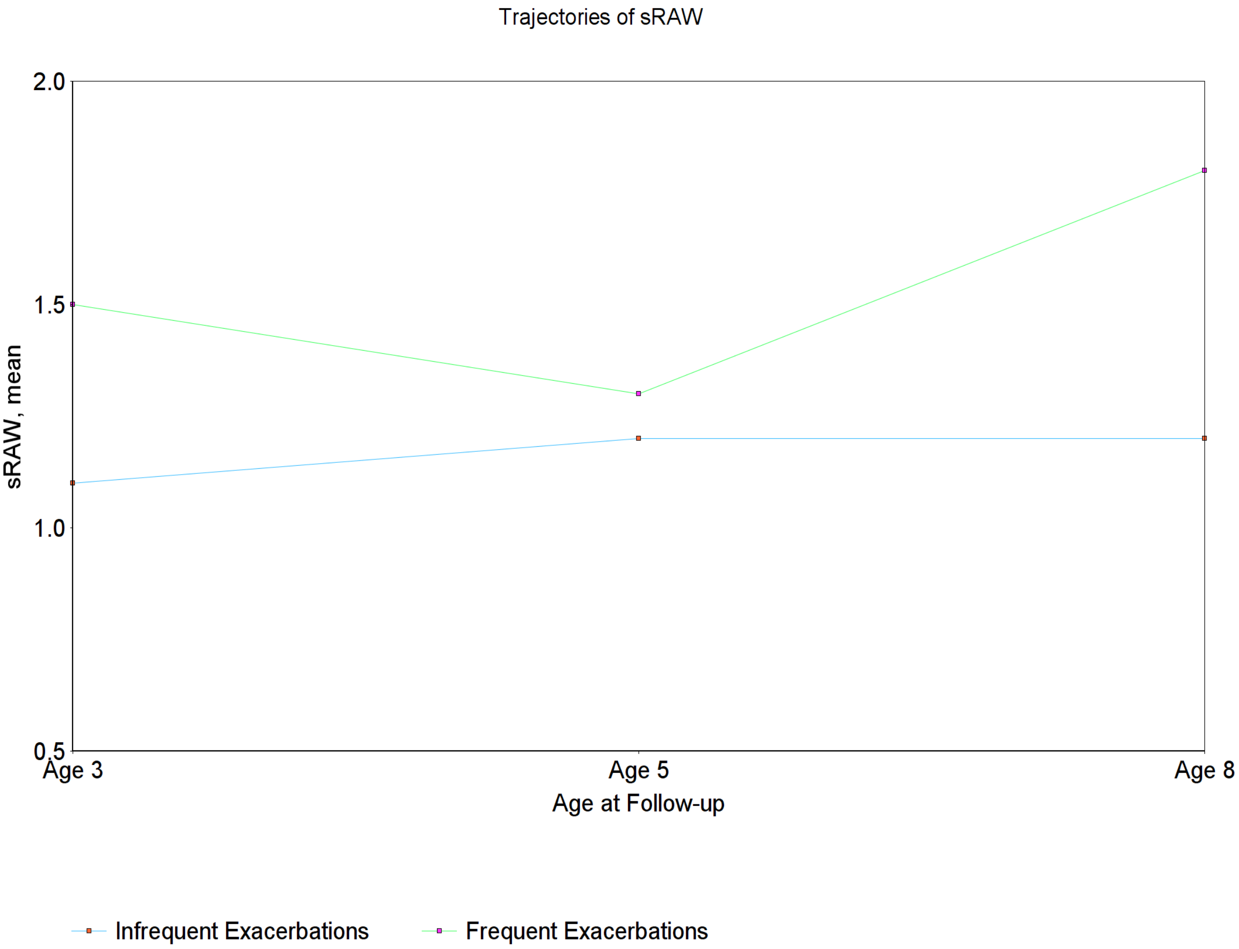


Figure E5.3: *Trajectories of airway inflammation from age 8 to age 16 years among children who never wheezed, those who wheeze but had no exacerbations, and two exacerbation clusters.*

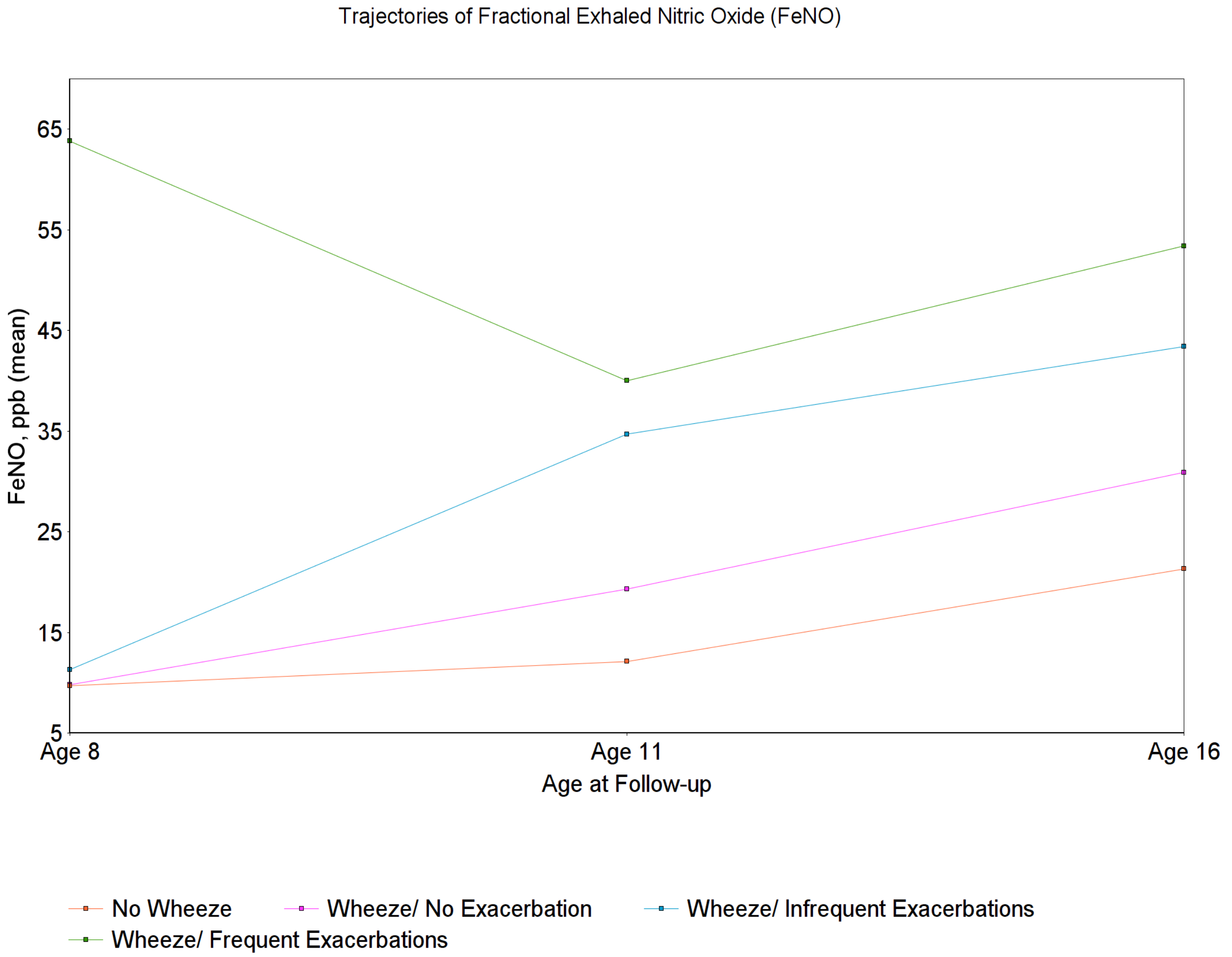
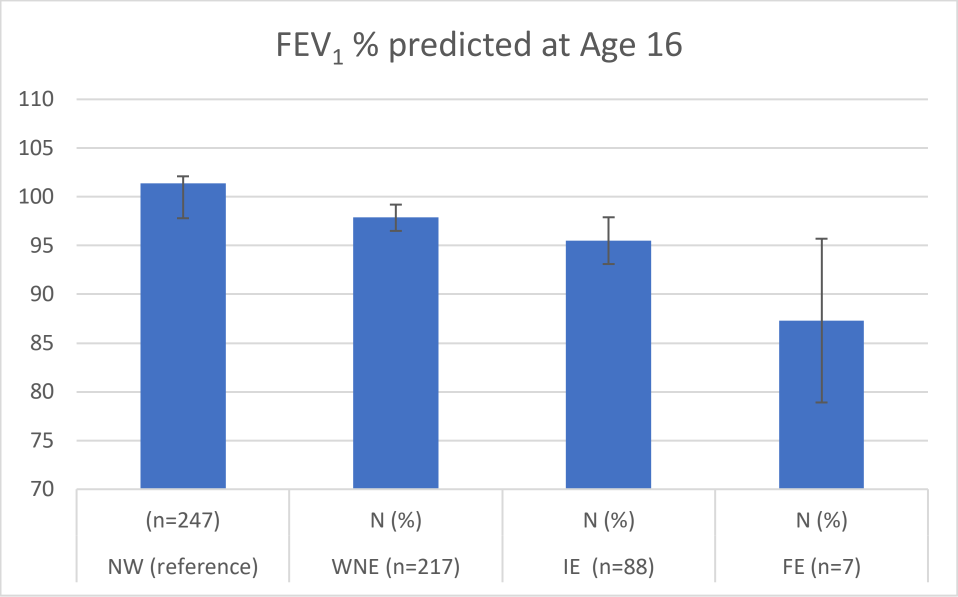
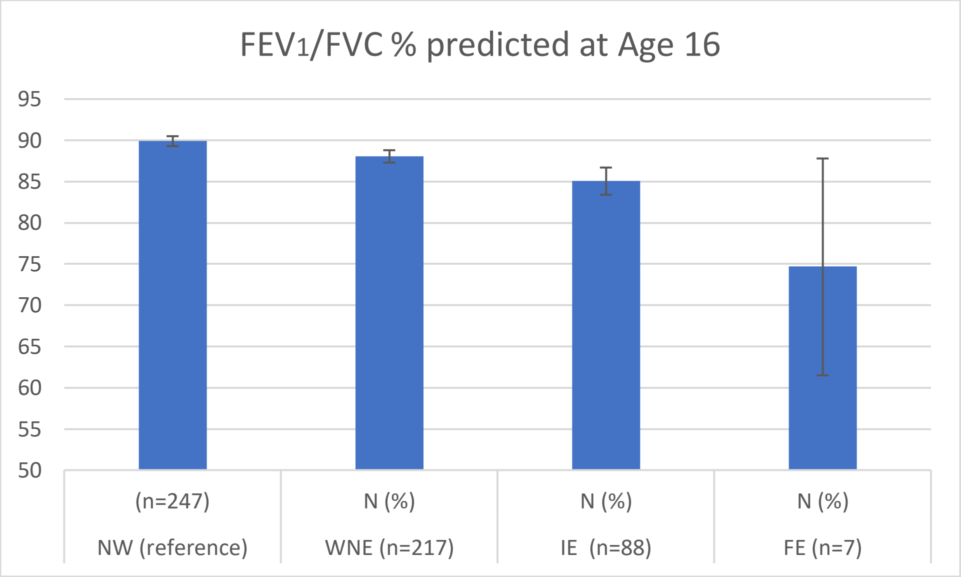


Figure E5.4: *Lung function at age 16 years among children who never wheezed (NW), those who wheezed, but have not had exacerbations (WNE), and children in the two exacerbation clusters (IE and FE). Children with lung function tests, N=559. FEV_1= forced expiratory volume in 1 second, FeNO= fraction of exhaled nitrogen oxide. Quantitative variables represented as mean (95% confidence interval).*

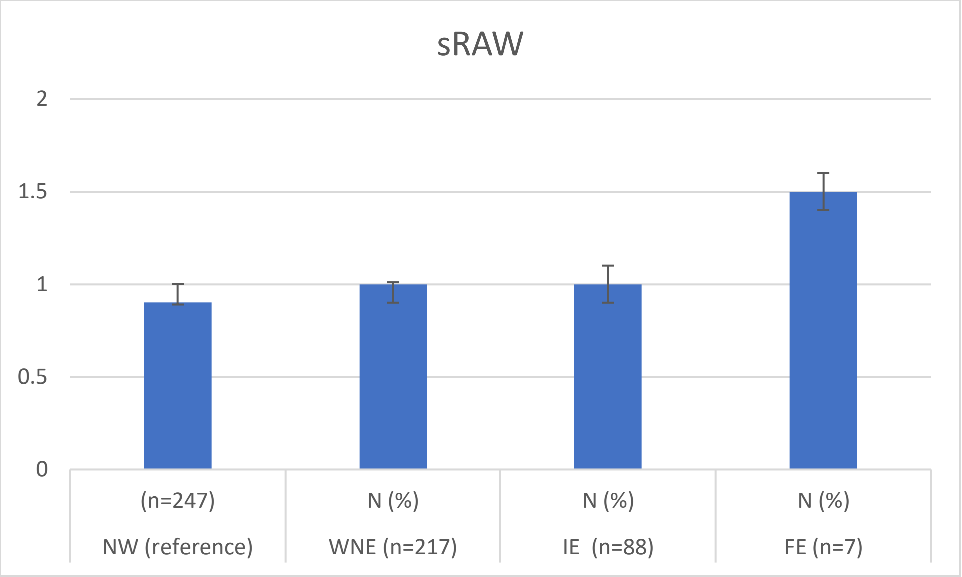
a) FEV_1 % predicted



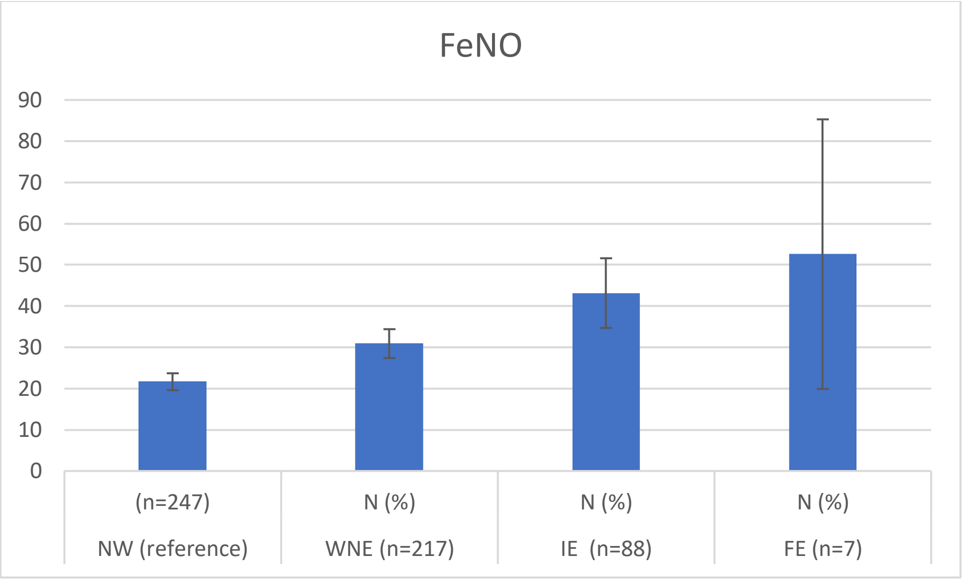
b) FEV_1/FVC (%)



c) **sRAW (kPa/s)**



d) **FeNO (ppb)**



5.8.3 References

1. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and wheezing in the first six years of life. The Group Health Medical Associates. *The New England journal of medicine* 1995; **332**(3): 133-8.

2. Lowe LA, Simpson A, Woodcock A, Morris J, Murray CS, Custovic A. Wheeze phenotypes and lung function in preschool children. *American journal of respiratory and critical care medicine* 2005; **171**(3): 231-7.

3. James DR, Lyttle MD. British guideline on the management of asthma: SIGN Clinical Guideline 141, 2014. *Arch Dis Child Educ Pract Ed* 2016; **101**(6): 319-22.

4. Genolini C, Falissard B. KmL: a package to cluster longitudinal data. *Comput Methods Programs Biomed* 2011; **104**(3): e112-21.

Blank page

Chapter 6 Patterns of wheeze severity from early childhood to late adolescence: Longitudinal transition analysis in a birth cohort study

Matea Deliu MD, Sara Fontanella PhD, Clare Murray MD, Angela Simpson MD PhD, Adnan Custovic MD PhD FAAAI

6.1. Rationale for this study

Characterising asthma severity has implications on guiding management and identifying children at risk of severe exacerbations. There is no universal consensus on how to define asthma severity, however the majority of national guidelines view asthma severity according to current treatments and so a stepwise management approach is employed.[1,2] Consequently, our initial approach to analysing patterns of asthma severity was taken from a treatment modality perspective using British Thoracic Society (BTS) guidelines.[2]

Using the Manchester Asthma and Allergy cohort, as in the previous chapter, we analysed 816 children from ages 3-16 who had data for at least three time points. Missing data was imputed using the k-nearest neighbour imputation algorithm.[3] We used the variable "BTS" which corresponded to the BTS treatment guidelines: 1) no asthma treatment, 2) step 1 (short acting bronchodilator), 3) step 2 (inhaled corticosteroid), 4) step 3 (add on long acting bronchodilator +/- leukotriene receptor blocker +/- increase inhaled corticosteroid). We did not have any children on higher steps of the treatment algorithm. Using latent class analysis and the BIC for model of best fit (Figure Intro 6.1), a 3 class model emerged. Figure Intro 6.2 shows the distribution of children per class: Class 1: n=635, class 2: n=104, class 3: n= 77. We labelled class 1: no asthma treatment, class 2: mild asthma, class 3: moderate asthma. Figure Intro 6.3 shows the trajectories of the classes. Children in class 1 remain on no asthma treatment throughout childhood. Children in class 2 consistently have mild asthma. Children in class 3 start having moderate asthma from age 5.

Figure Intro 6.1: *BIC model of best fit showed that a 3-class model best described the data.*

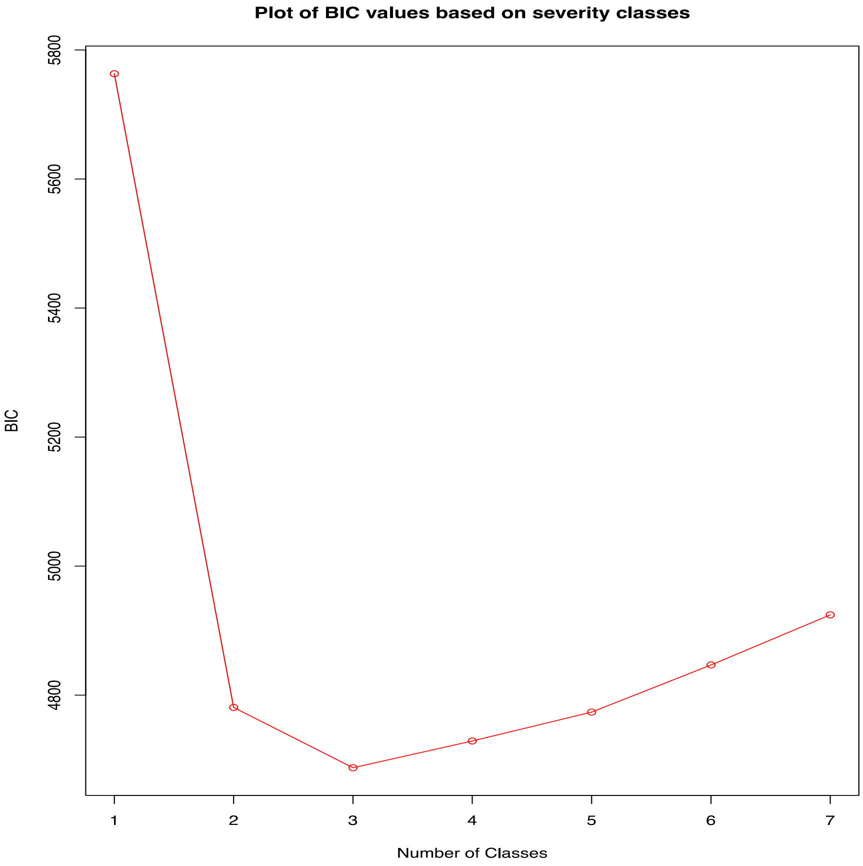
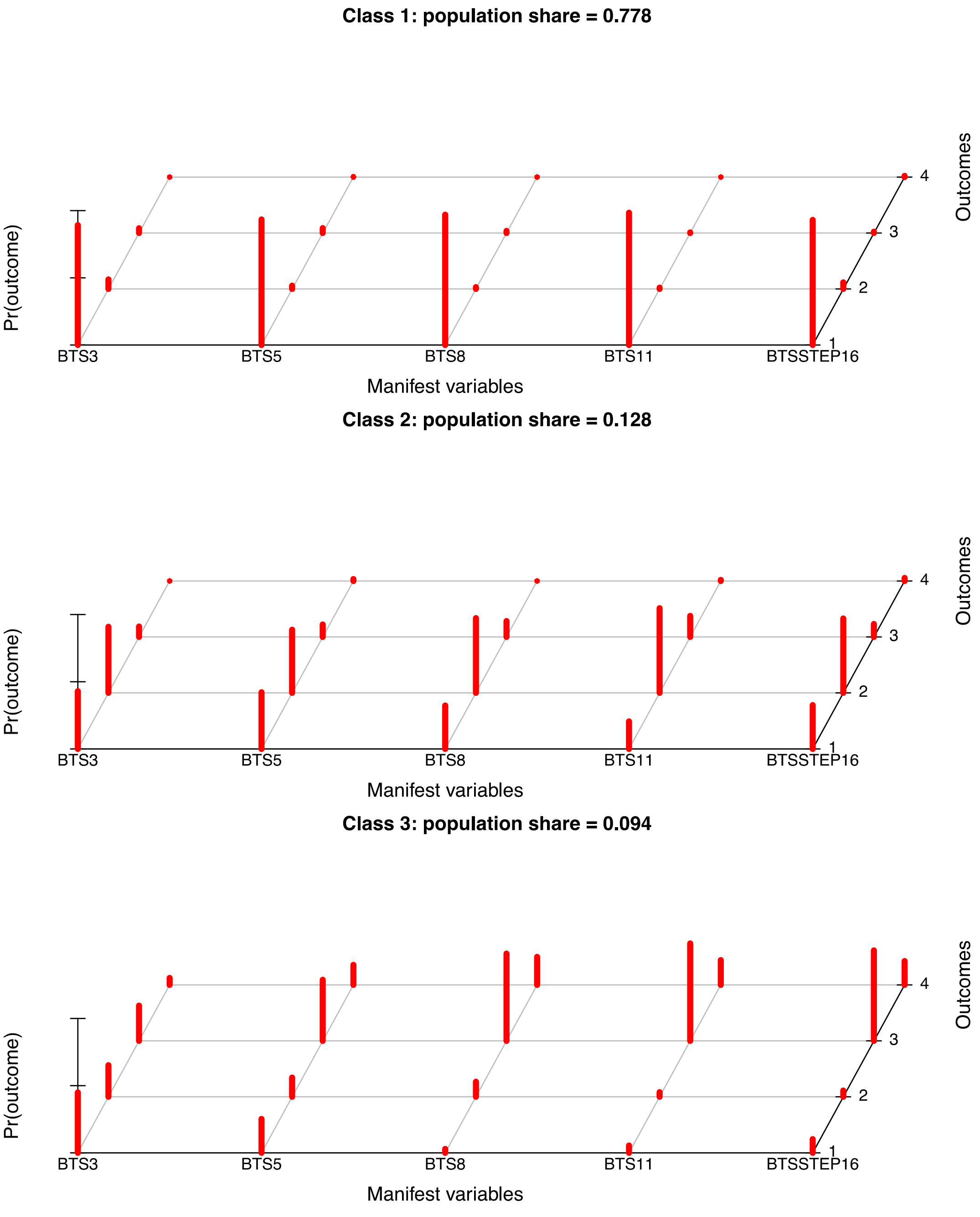


__Figure Intro 6.2__: Distribution of children per class and the probability of belonging to that class.

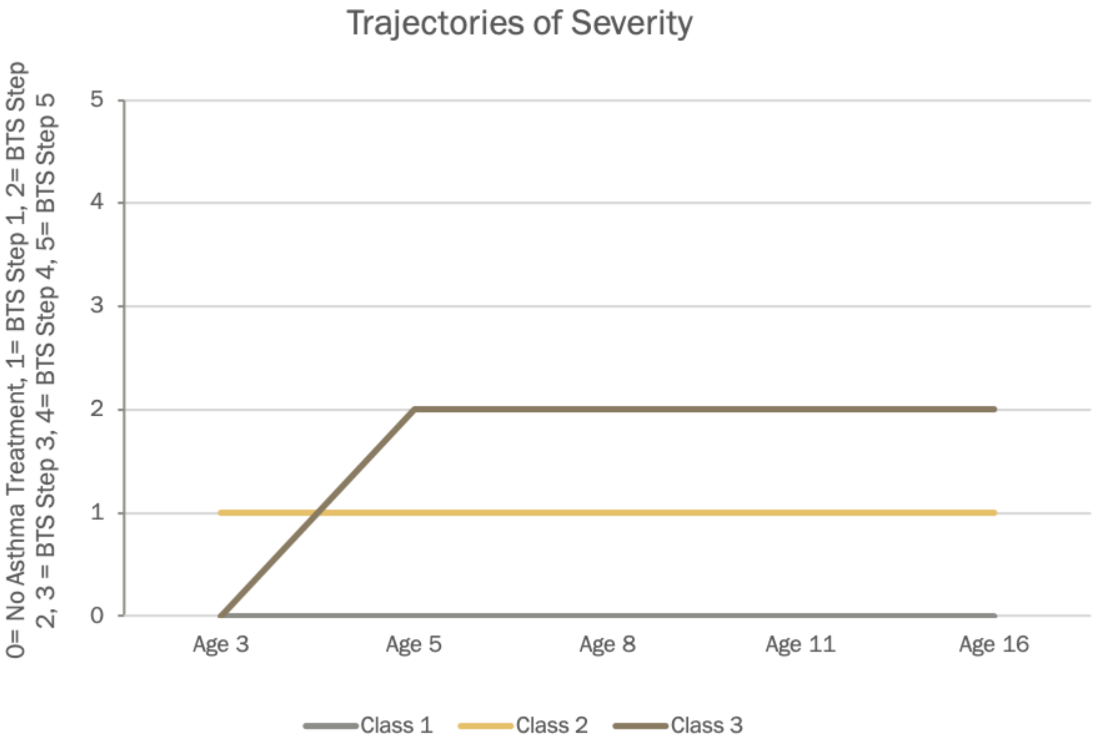


Class 1: population share = 0.778

Class 2: population share = 0.128

Class 3: population share = 0.094

Figure Intro 6.3: *Trajectories of asthma severity classes.*



The next three figures (Intro 6.4a-c) show the probability of belonging to each class. It is evident that children allocated to class 1 have a high probability of remaining in that class. In other words, children with no asthma treatment remain on no medication. Children in class 2 are a little more heterogeneous in that there is a large proportion of children that are on short acting bronchodilators, but also a sizeable proportion on no asthma treatment. Children in class 3 have a high probability of being on the BTS step 2 ladder by age 16, but also a smaller portion being on the BTS step 3 ladder starting from age 8 and continuing throughout childhood.

Figure Intro 6.4a: *Probability of belonging to class 1.*

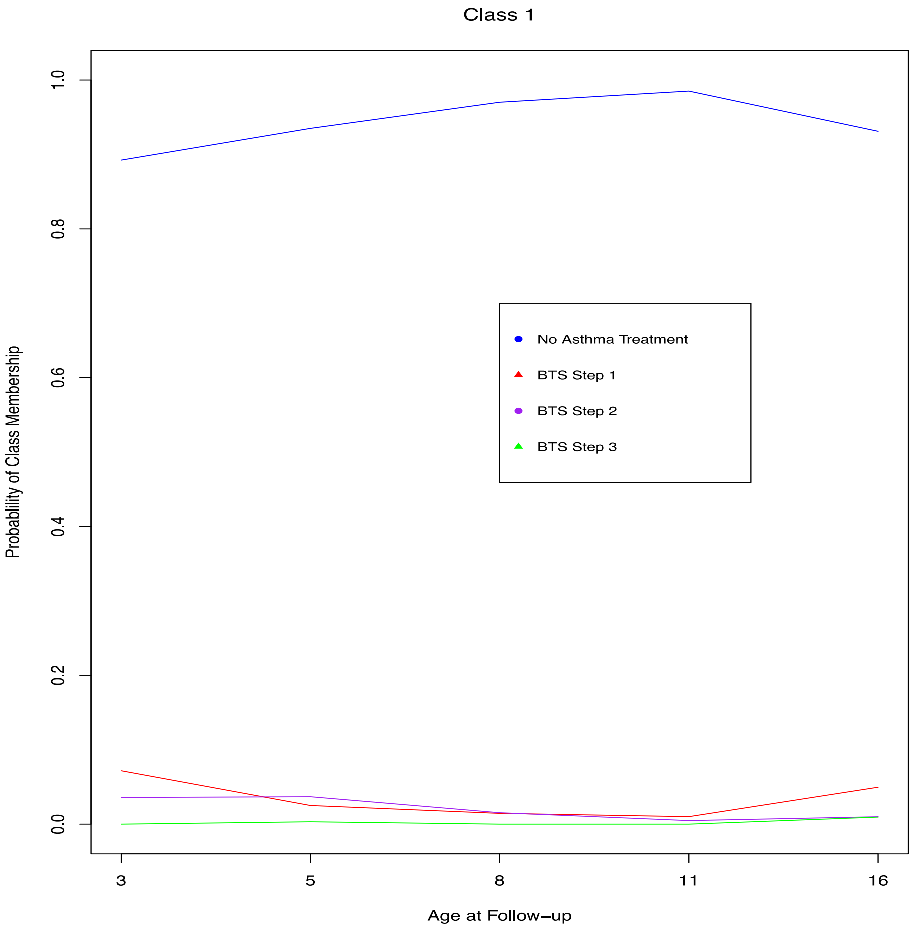


Figure Intro 6.4b: *Probability of belonging to class 2*

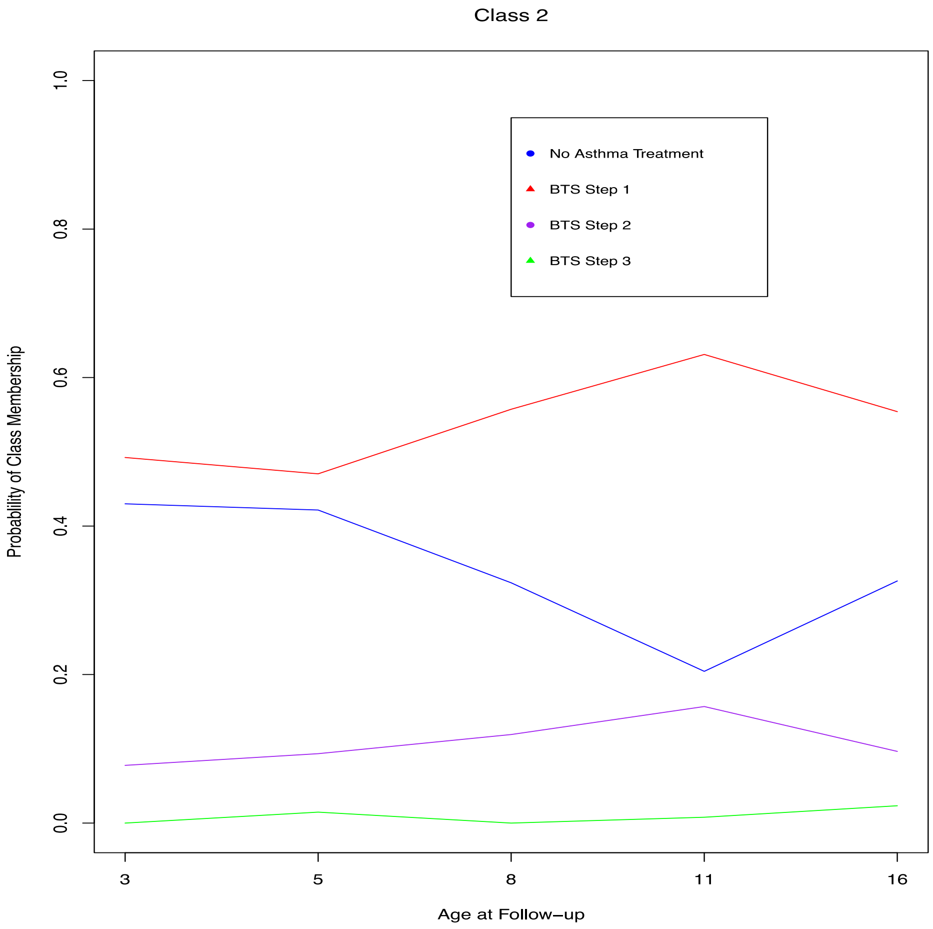
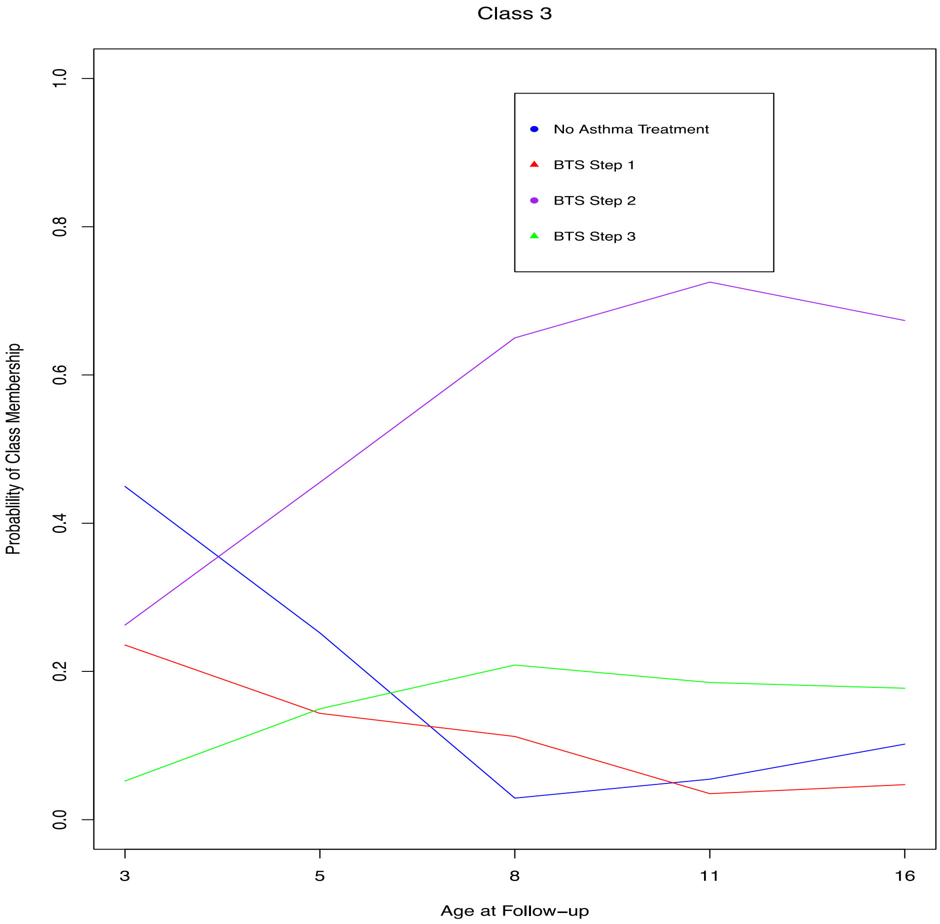


Figure Intro 6.4c: *Probability of belonging to class 3*



We then looked at the associations of the classes to lung function throughout childhood and essentially found no differences (Table Intro 6.1).

Table Intro 6.1: *Associations of severity classes with lung function: multinomial logistic regression using children who are not on treatment (Class 1) as the reference. FEV$_1$= forced expiratory volume in 1 second, FVC= forced vital capacity, FeNO= fraction of exhaled nitrogen oxide*

	Class 1, n=638 **Mean (SD)** **reference**	**Class 2, n=101** **Mean (SD)** **RR(95%CI)** **P-value**	**Class 3, n=77** **Mean (SD)** **RR(95%CI)** **P-value**
FEV1, mean (SD), age 5	96.6 (12.4)	95.6 (12.4) 0.99 (0.98-1.01) p=0.39	96.8 (12.9) 1.00 (0.98-1.02) p=0.88
FEV1, mean (SD), age 8	99.0 (11.8)	98.7 (14.3) 0.99 (0.97-1.01) p=0.87	98.7 (10.2) 0.99 (0.97-1.02) p=0.83
FEV, mean (SD), age 11	98.2 (11.5)	99.7 (12.4) 1.01 (0.99-1.03) p=0.27	99.1 (11.6) 1.00 (0.98-1.03) p=0.57
FEV, mean (SD), age 16	98.7 (12.2)	99.5 (11.9) 1.00 (0.98-1.03) p=0.65	100.3 (13.4) 1.01 (0.98-1.04) p=0.43
sRAW, mean (SD), age 3	1.1 (0.2)	1.1 (0.3) 1.06 (0.33-3.40) p=0.92	1.1 (0.3) 1.37 (0.41-4.60) p=0.61
sRAW, mean (SD), age 5	1.1 (0.2)	1.1 (0.2) 1.09 (0.43-2.78) p=0.84	1.2 (0.3) 1.37 (0.51-3.63) p=0.53
sRAW, mean (SD), age 8	1.2 (0.3)	1.2 (0.3) 0.78 (0.32-1.91) p=0.59	1.2 (0.3) 0.96 (0.39-2.34) p=0.93
sRAW, mean (SD), age 11	1.2 (0.3)	1.2 (0.4) 0.85 (0.40-1.79) p=0.67	1.3 (0.4) 1.43 (0.71-2.90) p=0.31
sRAW, mean (SD), age 16	0.9 (0.4)	0.9 (0.2) 0.90 (0.25-3.26) p=0.87	1.0 (0.2) 2.34 (0.54-10.07) p=0.25
Methacholine ddr slope, mean (SD), age 8	**72.8 (556.9)**	**58.5 (165.9)** **0.99 (0.98-1.00)** **p=0.85**	**149.3 (940.9)** **0.99 (0.98-1.00)** **p=0.42**
Methacholine ddr slope, mean (SD), age 11	128.9(581.6)	97.3 (188.4) 0.99 (0.98-1.00) p=0.76	190.4 (465.9) 1.00 (0.99-1.01) p=0.55

FEV1/FVC, mean (SD), age 8	86.7 (6.0)	86.4 (6.5) 1.00 (0.96-1.04) p=0.86	86.5 (5.8) 1.00 (0.96-1.05) p=0.74
FEV1/FVC, mean (SD), age 11	86.4 (7.3)	88.1 (5.2) 1.04 (1.00-1.08) p=0.04	86.6 (6.3) 1.00 (0.96-1.04) p=0.92
FEV1/FVC, mean (SD), age 16	87.8 (8.1)	89.1 (6.7) 1.02 (0.98-1.07) p=0.21	88.1 (5.2) 1.01 (0.96-1.05) p=0.81
FeNO, mean (SD), age 8	15.4 (17.1)	18.9 (18.3) 1.00 (0.99-1.03) p=0.26	24.7 (27.2) 1.02 (1.00-1.03) p=0.009
FeNO, mean (SD), age 11	18.6 (22.1)	20.1 (23.9) 1.00 (0.99-1.01) p=0.64	22.1 (9.9) 1.00 (0.99-1.02) p=0.30
FeNO, mean (SD), age 16	27.2 (30.2)	29.2 (29.6) 1.00 (0.99-1.01) p=0.69	29.5 (35.9) 1.00 (0.99-1.01) p=0.68

These results demonstrate that using current treatment steps is not a robust enough way of classifying asthma severity as it hasn't captured the heterogeneity of the disease. It isn't necessarily capturing symptom control and the level of treatment is independent of the level of symptoms. Indeed, severity is not a static feature of asthma but rather a dynamic process that changes with time (as will be demonstrated with the next analysis). Furthermore, responsiveness to treatment is not uniform, even among patients with similar severity.[4-7] Therefore, the use of severity as a single outcome measure has limited value in ascertaining which treatment is required and predicting its response.[8] As a result, a different approach was taken focusing on severity of symptoms.

6.1.1 References

1. Reddel HK, Bateman ED, Becker A, et al. A summary of the new GINA strategy: a roadmap to asthma control. *Eur Respir J* 2015; **46**(3): 622-39.
2. Network BTSSIG. British guideline on the management of asthma: A national clinical guideline. *Thorax* 2014; (69): 1-192.
3. P. Jonsson CW. An evaluation of k-nearest neighbour imputation using Likert data. 10th International Symposium on Software Metrics, 2004 Proceedings; 2004; Chicago, Illinois, USA, USA: IEEE; 2004.
4. Fitzpatrick AM. Severe Asthma in Children: Lessons Learned and Future Directions. *J Allergy Clin Immunol Pract* 2016; **4**(1): 11-9; quiz 20-1.
5. Chung KF. Precision medicine in asthma: linking phenotypes to targeted treatments. *Curr Opin Pulm Med* 2018; **24**(1): 4-10.
6. Selby L, Saglani S. Severe asthma in children: therapeutic considerations. *Curr Opin Allergy Clin Immunol* 2019; **19**(2): 132-40.
7. Bush A, Saglani S, Fleming L. Severe asthma: looking beyond the amount of medication. *Lancet Respir Med* 2017; **5**(11): 844-6.
8. Stoloff SW, Boushey HA. Severity, control, and responsiveness in asthma. *J Allergy Clin Immunol* 2006; **117**(3): 544-8.

6.2 Abstract

Background: Wheeze phenotypes are dynamic and children can transition between different phenotypes and severity states.

Aim: To investigate patterns and wheezing severity from early childhood to late adolescence.

Methods: In an unselected birth cohort, we applied a longitudinal latent transition (Markov) model to ascertain patterns of wheezing severity throughout childhood.

Results: The optimal solution in a multivariate latent Markov model was a 3-state model with homogeneous transition probabilities: a healthy-no symptom state, and two states of severity and: mild/moderate wheeze, and severe wheeze. Children with severe wheeze tend to remain in the severe wheeze state. Children with mild wheeze tend to transition frequently throughout childhood particularly in the no wheeze/healthy state. Children with severe wheeze are more likely to be atopic, have greater airway resistance and lung inflammation, and poorer lung function by late childhood when compared to children with mild/wheeze and those that transition frequently between wheeze states. Furthermore, children assigned to a persistent wheeze phenotype (persistent troublesome and persistent controlled) have a low tendency of transitioning into healthier, or less severe states.

Conclusion: Utilizing a data-driven method, we have shown that presence or absence of wheeze is not a robust enough feature to ascertain phenotypes of disease. Furthermore, wheeze phenotypes are not static but rather dynamic processes involving children transitioning between states of no symptoms to mild/moderate and severe symptoms. This provides a framework for identifying the heterogeneity of disease severity occurring at a population level.

6.3 Introduction

Wheeze is amongst the most common symptoms in childhood, with parents of almost a third of children reporting that their child has wheezed on at least one occasion before age 3[1,2]. Wheezing in childhood is highly heterogeneous,[1,3-5] and although the majority of children stop wheezing by school age, a proportion of them have wheeze persisting into late childhood and adulthood[6]. However, it is difficult to predict which of these children will stop and which will persist or develop asthma[7].

The seminal work in identifying wheeze phenotypes from Martinez et al used clinical observations to identify three ,mutually exclusive wheeze phenotypes: transient early, late-onset, and persistent[1]. This work was then expanded by various research groups using data-driven methodologies such as the latent class analysis (LCA), which assigns individuals to latent classes based on their membership probability[1,3,4,8,9]. Applying this methodology to the presence or absence of current wheezing over time, a further one or two intermediate phenotypes have been identified[4,10], and incorporation of the data from health care records has divided persistent wheeze into persistent controlled and persistent troublesome classes (of which persistent troublesome wheeze was associated with poor lung function, and severe exacerbations despite treatment)[3]. However, a recent study which pooled data from five birth cohorts has shown that more than 10% of the population cannot be assigned to a specific wheeze class with high certainty ascertaining, that there are individuals whose patterns of wheezing do not fit well within the assigned phenotypes (particularly within the late-onset class), and that a notable proportion of transient wheezers have asthma in later life, thereby concluding that presence or absence of wheeze is unlikely to be robust enough to deriving internally homogenous phenotypes[10]. Incorporating information on the severity of wheezing may offer additional valuable information to help our understanding of the heterogeneity and progression of childhood wheezing and asthma. However, there is no consensus on how to define asthma severity, and the majority of national guidelines view asthma severity according to current treatments requirements[11,12]. It is likely that the presence and severity of respiratory symptoms are not static features of asthma but rather dynamic processes that changes with time within individual patients, and that individuals may transitions over time from one state to another.

While LCA has been used extensively to derive wheeze phenotypes, few studies have explored how children transition during childhood. Latent transition analysis (LTA) has been used in the Isle of Wight cohort to simultaneously identify classes of asthma and wheezing and characterize transition probabilities over time[13], and in a study by Garden *et al* to incorporate several asthma domains and show that transition between asthma phenotypes was common in early childhood but less common in later childhood[14]. We hypothesise that wheeze phenotypes are not static during childhood, but rather a dynamic process involving individual children transitioning into different states of symptoms presence and severity, and that these transitions may be an important feature of the disease. To address our hypotheses, we used LTA methodology to ascertain the heterogeneity of wheezing patterns and severity from early childhood to late adolescence. We also sought to examine how children within different wheeze phenotypes determined using LCA transition between states of symptom expression and states of control/severity.

6.4 Methods

6.4.1 Study Population and data source

The Manchester Asthma and Allergy Study is a population-based birth cohort.[15,16] Subjects were recruited prenatally and followed prospectively to age 16 years. The study was approved by the Local Ethics Committee. Parents provided written informed consent.

Children attended clinical follow-up at ages 1, 3, 5, 8, 11, and 16 years. Validated questionnaires were interviewer administered and parents provided information on symptoms, treatments received, and environmental exposures. If study participants were unable to attend clinic, home visits were carried out.

6.4.2 Definition of variables

Current wheeze: Parentally reported positive answer to the question: "Has your child had wheezing of whistling in the chest in the past 12 months"?

Asthma: At least two of the following three features: 1) current wheeze; 2) current use of asthma medication; 3) physician-diagnosed asthma ever (flexible criterion). We also utilized a strict criterion (positive answer to all three features). Data was available from ages 3-16.

6.4.2.1 Variables used to ascertain patterns severity of wheeze/asthma

Wheeze attacks: Parentally reported quantifiable answer to the question: "How many attacks of wheezing has your child had in the past 12 months? None, 1-4, more than 4."

Sleep disturbance: Parentally reported quantifiable answer to the question: "In the past 12 months, how often, on average has your child's sleep been disturbed due to wheezing? Never, 1 night per week, more than 1 night per week.

Speech limitation: Parentally reported positive answer to the question: “In the past 12 months, has wheezing been severe enough to limit your child’s speech to only one or two words at a time between breaths?”

Exercise induced wheeze: Parentally reported positive answer to the question: “In the past 12 months has your child’s chest sounded wheezy during or after exercise?”

We defined severe wheeze as having all four symptoms at the most severe spectrum. These variables were available from ages 5-16 years.

6.4.2.2 Wheeze phenotypes

Using a longitudinal latent class item response model, children were assigned to 5 classes[3]: (1) No wheezing (NW), (2) Transient early wheezing (TEW), (3) Late-onset wheezing (LOW), (4) Persistent controlled wheezing (PCW) and (5) persistent troublesome wheezing (PTW).

6.4.2.3 Variables used to test the validity of severity patterns

At age 16 years, we measured specific airway resistance (sRaw) using plethysmography at ages. FEV_1 and forced vital capacity (FVC) were measured by using spirometry; we recorded percent predicted FEV_1 and the FEV1/FVC ratio. Airway hyperreactivity (AHR) was assessed in a 5-step protocol by using quadrupling doses of methacholine; children were categorized as having AHR after a 20% decrease in FEV1 by the final stage of the challenge (16 mg/mL). We calculated the dose-response slope16 to include all evaluable data as a continuous variable. Atopic sensitization was ascertained by using skin prick tests (SPTs) to a panel of inhalant and food allergens (details can be found in the Methods section in this article's Online Repository);

we defined atopy as a wheal 3 mm larger than that elicited by the negative control to at least 1 allergen.

6.4.3 Statistical Analysis

Latent Markov Models to investigate wheeze severity

We applied Latent Markov (LM) models[17] to define the evolution of wheeze and its severity using the following variables: current wheeze (binary), number wheeze attacks (categorical), wheeze disturbing sleep (categorical), wheeze limiting speech (binary), and exercise induced wheeze (binary). LM models consider the analysis of longitudinal data when the response variables measure a phenomenon of interest that is not directly observable. The characteristic of interest and its evolution in time are described by a latent process that is assumed to follow a Markov chain with a certain number of states, typically referred to as latent states, and, given this process, the response variables are assumed to be conditionally independent. The basic idea related to this assumption, which is referred to as *local independence*, is that the latent process fully explains the observable behaviour of a subject[17]. Model selection was performed using the Bayesian Information Criterion (BIC) index. Statistical analyses were performed in R through the *LMest* package[18].

6.5 Results

6.5.1 Participant Flow and Demographic Data

1184 children born into the cohort had some evaluable data. Of those, we excluded 133 children who were randomised into the environmental control arm[19]. From the remaining children, 545 had complete data on parentally reported current wheeze and severity symptoms at all four follow-up visits. Demographic characteristics were similar between children with and without complete data (Table 6.1 and E6.1). Current wheeze was present in 14-20% of children at any one time during the follow-up. At least a third of our population had comorbid eczema and/or rhinitis.

Table 6.1: *Demographics of our study population, n=545, with complete dataset*

Boys (whole population) Girls (whole population)	290, 53% 255, 47%				
Family history of asthma	149, 27%				
Family history of atopy	444, 81%				
	Age 3	**Age 5**	**Age 8**	**Age 11**	**Age 16**
BMI, median (IQR)	16.6 (15.7-17.4)	16.2 (15.3-17.1)	16.5 (15.4-17.9)	18.3 (16.7-20.5)	21.3 (19.7-23.6)
Maternal smoking	72, 13%	75, 14%	74, 14%	66, 12%	64, 12%
Paternal smoking	106, 19%	104, 19%	97, 18%	71, 13%	64, 12%
Asthma ever	118, 22%	197, 36%	150, 28%	153, 28%	79, 14%
Atopic sensitization (SPT)	111, 20%	138, 25%	158, 29%	164, 30%	235, 43%
Current rhinitis	n/a	147, 27%	161, 29%	186, 34%	216, 40%
Current eczema	142, 26%	179, 33%	124, 23%	102, 18%	99, 18%
Any current asthma medication	58, 11%	91, 17%	81, 15%	92, 17%	74, 14%
Current wheeze, parentally reported	110, 20%	101, 18%	91, 17%	83, 15%	78, 14%
Current asthma, parentally reported	63, 12%	89, 16%	82, 15%	92, 17%	80, 15%
	Asthma severity symptoms				
Number of wheeze attacks					
none	n/a	445, 81%	454, 83%	462, 85%	468, 86%
1-3	n/a	69, 13%	63, 12%	49, 9%	52, 10%
4 or more	n/a	31, 6%	28, 5%	34, 6%	25, 4%
Exercise induced wheeze		35, 6%	36, 7%	49, 9%	45, 8%
Wheeze affecting sleep					
none	n/a	493, 91%	491, 90%	505, 93%	518, 95%
<1 night per week	n/a	28, 5%	36, 7%	27, 9%	18, 3%
More than 1 night per week	n/a	24, 4%	18, 3%	13, 2%	9, 2%
Wheeze limiting speech	n/a	19, 3%	16, 3%	79, 14%	9, 2%
Children considered to be severe	n/a	18, 3%	14, 2.5%	13, 2%	7, 1.3%

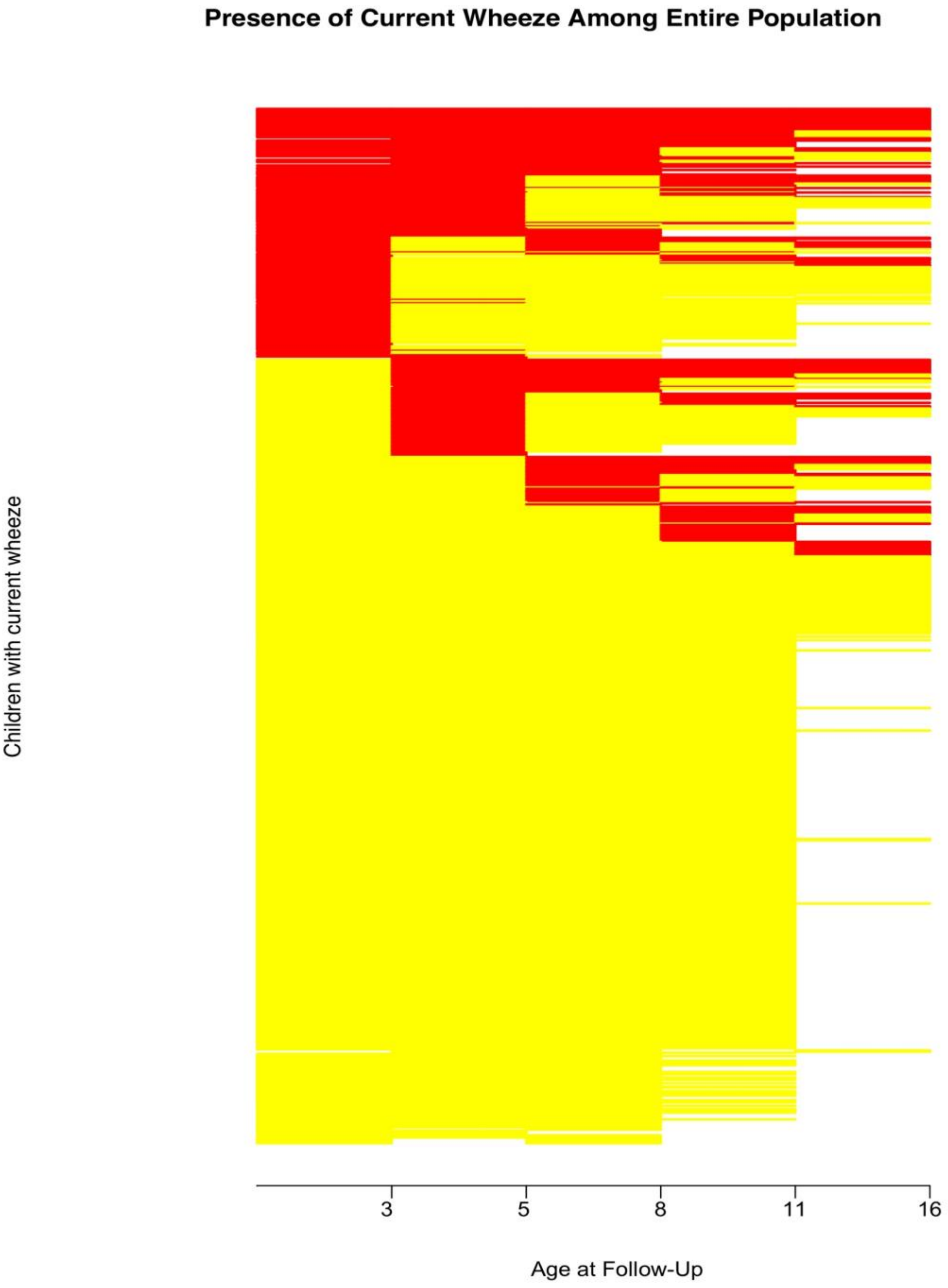
6.5.2 Descriptive statistics: Current wheeze and asthma from early childhood to adolescence

We first mapped the presence/absence of current wheeze and asthma during childhood in the whole study population (n=1051), and presented it as a heatmap in Figure 6.1. Children transitioned into and out of having wheeze at various stages throughout childhood (Figure 6.1a). The majority of wheeze occurred early on by age 3, and only a proportion of children (15.3%) continue to persistently wheeze throughout childhood. Using a flexible criterion for asthma definition resulted in almost identical heatmap as that for current wheeze (Figure 6.1b). As expected, the strict criterion reduced the number of children with asthma diagnosis but also resulted in a more uniform pattern throughout childhood (Figure 6.1c).

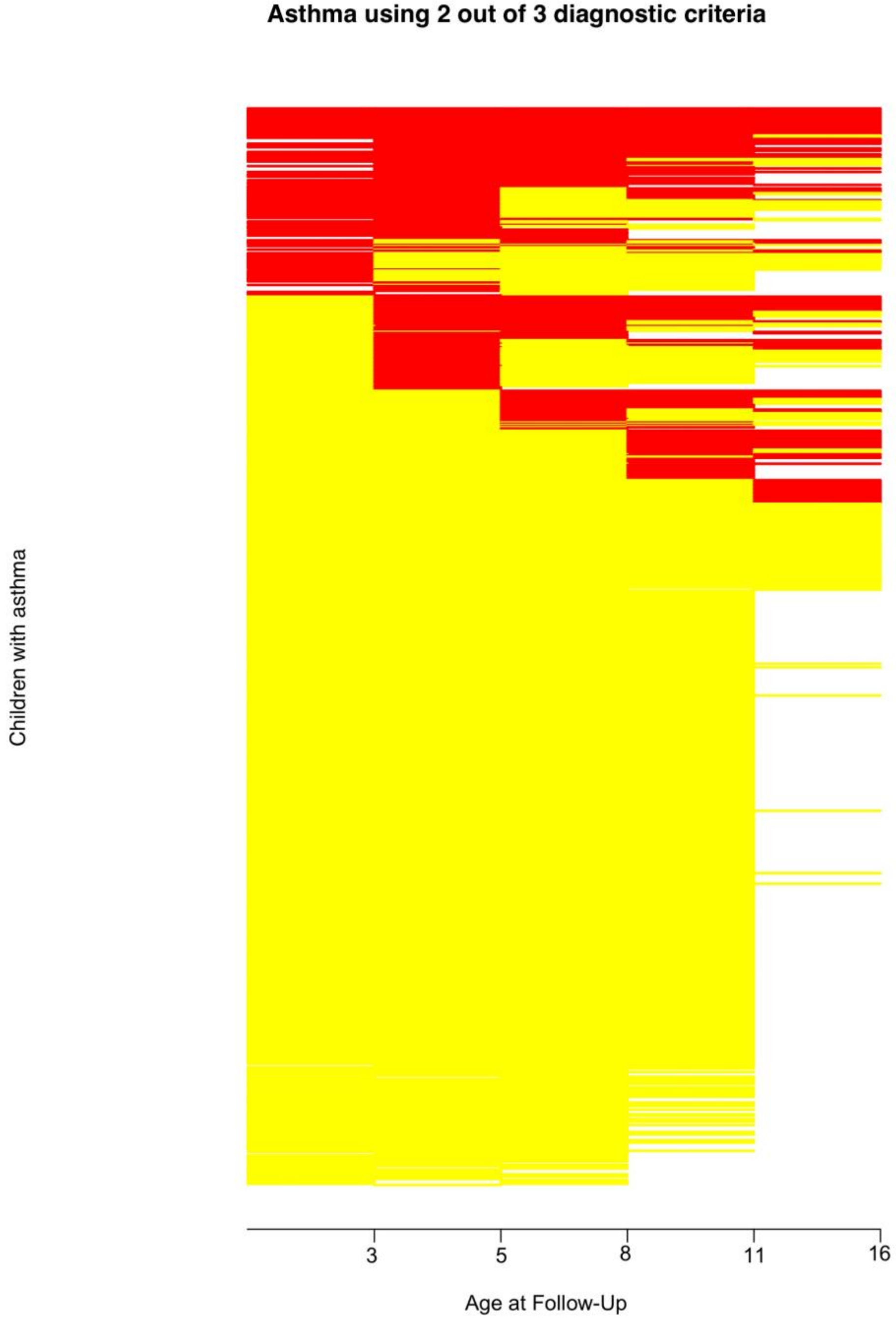
We then proceeded to investigate the longitudinal patterns of wheeze severity among 545 children with a complete data set. Of these, 151 (28%) experienced current wheeze at some point throughout childhood, of whom 18 (3.3%) had wheeze at every time point throughout the follow-up period.

Figure 6.1: *A heatmap showing the presence of current wheeze (a), current asthma (b) and current asthma (strict criteria) (c) among our entire population from age 3-16, N=1051.Each line represents a child. White spaces indicate missing response.*

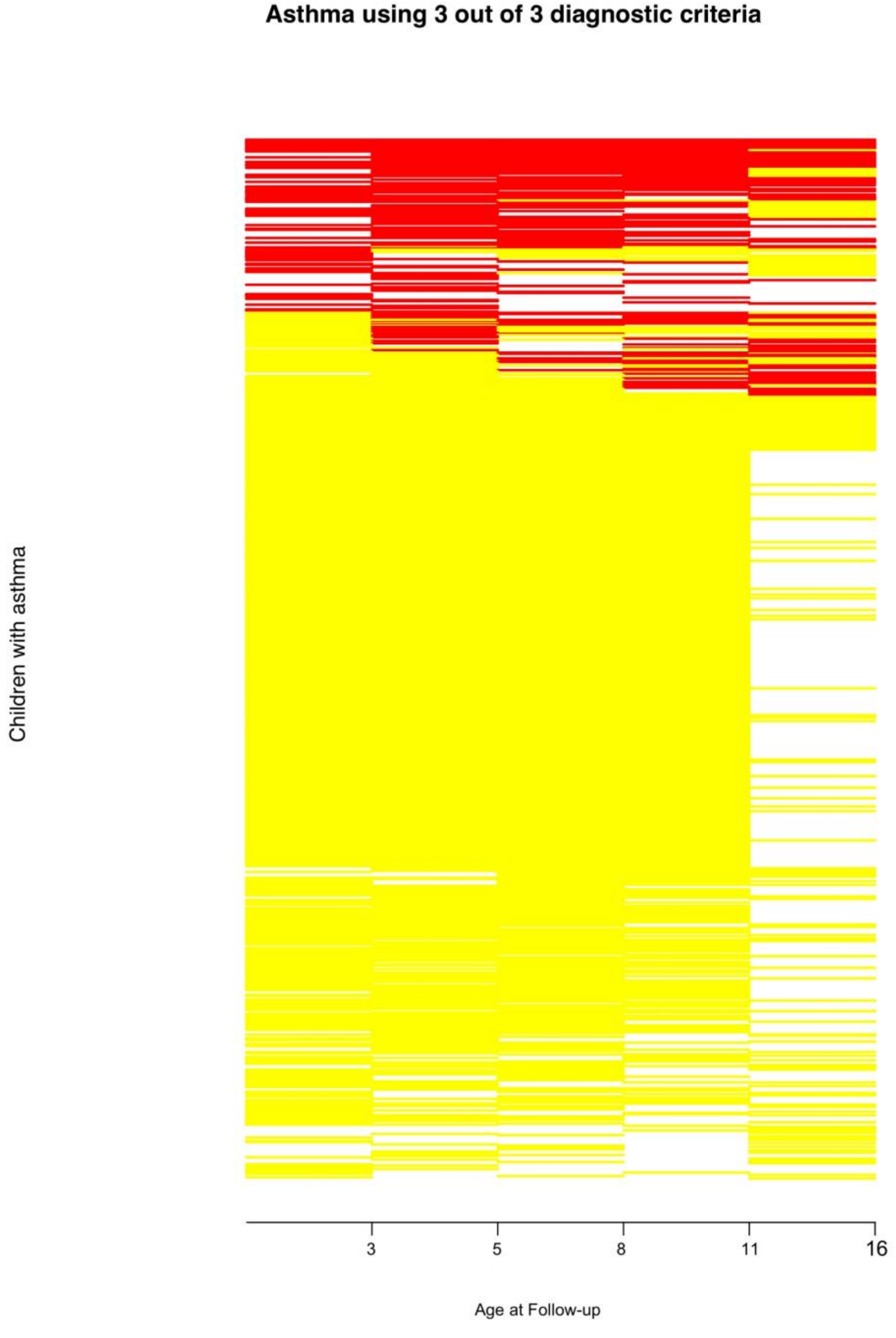
a)



b)


Asthma using 2 out of 3 diagnostic criteria
Children with asthma
3
5
8
11
16
Age at Follow-Up

c)


Asthma using 3 out of 3 diagnostic criteria
Children with asthma
3
5
8
11
16
Age at Follow-up

6.5.3 Latent Markov Models to investigate wheeze severity states throughout follow-up

We adopted a multivariate basic LM model, and the BIC indicated that the optimal solution was a 3-state model with homogeneous transition probabilities (Table E6.2). The estimates of the conditional response probabilities, shown in Table 6.2, suggested the following interpretation of the states: State 1, Mild/moderate wheeze (characterised by 1-3 wheeze attacks, 1 night per week of sleep disturbance due to wheeze, and some exercise induced wheeze); State 2, Severe wheeze (characterized by the presence of all symptoms and by a large number of subjects presenting at the most severe spectrum (>4 wheeze attacks, frequent exercise induced wheeze, >1 night per week of sleep disturbance due to wheeze); and State 3, healthy/no wheeze state, characterised by the absence of any of the considered symptoms.

At the beginning of the study, most (82%) of the subjects belong to the healthy latent state, whereas a small percentage (5.5%) of subjects belong to state 2 (severe wheeze) (Table E6.3). Figure 6.2 shows that 2% of children (n=10) remained in the severe wheeze state throughout the follow up period. Table 6.3 shows the descriptive characteristics of children who consistently remained in the different states (No Wheeze, n=366; Mild/Moderate, n=5; and Severe wheeze state, n=10) along with 162 children who transitioned frequently throughout the follow-up period. The highest prevalence of comorbidities (eczema and rhinitis, 40% and 90% respectively) was seen among the severe wheeze children. 100% and 70% of children who were sensitised to any aeroallergen were from the mild/moderate and severe states, respectively.

Transition probabilities (Table E6.4) show that children with no wheeze in the State 3 (healthy latent state) tend to remain in this state throughout childhood. Similarly, children with Severe wheeze (State 2) tend to remain in the state of severe wheeze, but to a lesser extent.

Table 6.2: *Estimates of the conditional response probabilities*

		States		
		1	**2**	**3**
Current wheeze	0	0.000	0.000	1.000
	1	1.000	1.000	0.000
Number of wheeze attacks: 0, 1-3, >4	0	0.006	0.000	1.000
	1	0.807	0.344	0.000
	2	0.187	0.656	0.000
Exercise induced wheeze	0	0.645	0.286	1.000
	1	0.355	0.714	0.000
Wheeze disturbing sleep: never, 1 night/week, >1 night/week	0	0.672	0.155	1.000
	1	0.236	0.469	0.000
	2	0.092	0.376	0.000
Wheeze limiting speech	0	0.899	0.508	0.976
	1	0.101	0.492	0.024

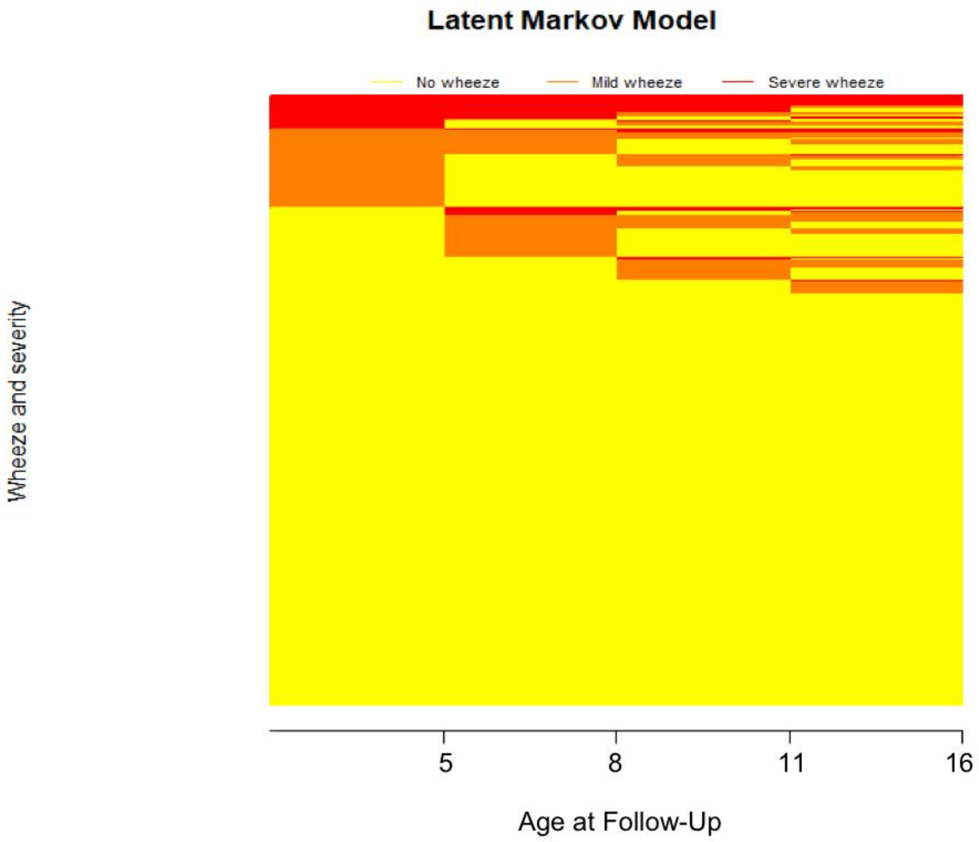
Table 6.3: *Descriptive table of children who consistently remained in the no wheeze state, the mild/moderate wheeze state, the severe state, as well as those that were frequently transitioning between states. Data taken at age 16.*

	No Wheeze (n=366)	Mild/Moderate (n=5)	Severe wheeze state (n=10)	Children who transitioned between states (n=162)
Boys (whole population) Girls (whole population)	184, 50.2% 182, 49.7%	4, 80% 1, 20%	5, 50% 5, 50%	90, 56% 72, 44%
Maternal smoking	106, 29%	4, 80%	7, 70%	81, 50%
Paternal smoking	19, 5%	1, 20%	2, 20%	19, 12%
Presence of asthma ever	37, 10%	5, 100%	9, 90%	87, 54%
Atopic sensitization (SPT)	134, 37%	5, 100%	7, 70%	89, 55%
Current rhinitis	118, 32%	2, 40%	9, 90%	87, 54%
Current eczema	52, 14%	1, 20%	4, 40%	42, 26%

6.5.4 Severity states and their associations with sensitisation and lung function

Using a multinomial logistic regression model (reference category children with no wheeze), we proceeded to analyse the associations of sensitisation, lung function and airway inflammation with the different wheeze states (Table 6.4). Children with severe wheeze and those children who transitioned frequently between states were significantly more likely to be sensitized, RR (95%CI): severe wheeze, 9.5 (1.2-78.5) p=0.03, children who transition, 2.3 (1.5-3.4) p<0.001. Children with severe wheeze were significantly more likely to have poorer lung function, had significantly higher FeNO, and greater airway resistance when compared to children with mild/moderate wheeze and children who transitioned between wheeze states (Table 4).

Figure 6.2: Estimated posterior distribution of the latent states for each subject and each time occasion stratified by wheeze severity. Each line represents a child.



6.5.5 Severity of clinical symptoms across latent wheeze phenotypes

We then looked at the latent state transitions between wheeze classes derived using LCA[3]. Figure 6.3 shows the transition of the severity of wheeze symptoms of each individual among the wheeze classes throughout childhood. Severe wheeze state was primarily present in the persistent controlled, persistent troublesome, and late onset wheeze class. About a quarter (23%) of children in the persistent troublesome wheeze class and 9% in the persistent controlled wheeze class remained in the severe wheeze state throughout the follow-up period. The majority of severe wheeze in the persistent controlled class occurred by age 3. Although the results indicated a lower tendency of transitioning to the healthy state in the PCW and PTW phenotypes, there was high within-class variability. The majority of children in transient early wheeze had no symptoms until age 8 after which a few children transition into expressing mild symptoms. Children associated only with the healthy state (state 3) were found in both TEW and LOW phenotypes.

Table 6.4: *Mutinomial logistic regression using wheeze severity states and asthma outcomes, lung function, sensitisation, airway resistance, and level of lung inflammation at age 16 years. No wheeze as reference.*

	No Wheeze (n=366) – reference Mean (95% CI) RR (95% CI)	Mild/Moderate Wheeze (n=5) Mean (95% CI) RR (95% CI)	P -value	Severe Wheeze (n=10) Mean (95% CI) RR (95% CI)	P value	Children who transitioned between states (n=162) Mean (95% CI) RR (95% CI)	P-value
Atopic sensitization (SPT) (n, %)	37, 10% N/a reference	5, 100% -	0.7	9, 90% 9.5 (1.2-78.5)	**0.03**	87, 54% 2.3 (1.5-3.4)	**<0.001**
Presence of asthma ever (n %)	134, 37% N/a reference	5, 100% 169 (-)	0.80	7, 70% 630 (-)	0.82	89, 55% 3.5 (0.6-21.9)	0.18
FEV_1/FVC %pred	89.3 (88.6-89.9) N/a reference	83.2 (72.4-93.9) 0.92 (0.8-0.99)	**0.03**	83.0 (79.1-86.9) 1.1 (0.8-1.25)	**0.008**	86.8 (85.5-88.2) 0.95 (0.92-0.98)	**0.001**
FeNO, ppb	21.2 (19.5-23.0) N/a reference	55.0 (17.5-127.5) 1.02 (1.00-1.04)	**0.007**	56.7 (18.5-95.0) 1.03 (1.01-1.04)	**<0.001**	34.5 (28.3-40.9) 1.01 (1.00-1.02)	**<0.001**
sRAW	0.95 (0.91-0.96) N/a reference	0.98 (0.5-1.4) 3.5 (0.03-183)	0.67	1.14 (0.94-1.34) 30.9 (1.5-602)	**0.02**	1.06 (1.01-1.09) 8.7 (3.3-22.8)	**<0.001**

Quantitative variables are represented as mean (95%CI); ordinal variables are represented as frequencies (%). RR relative risk, CI, confidence interval. Bolded values represent significant p-values. (-) signifies numbers too big to quantify.

Latent Markov Models to investigate severity of clinical symptoms amongst wheeze classes

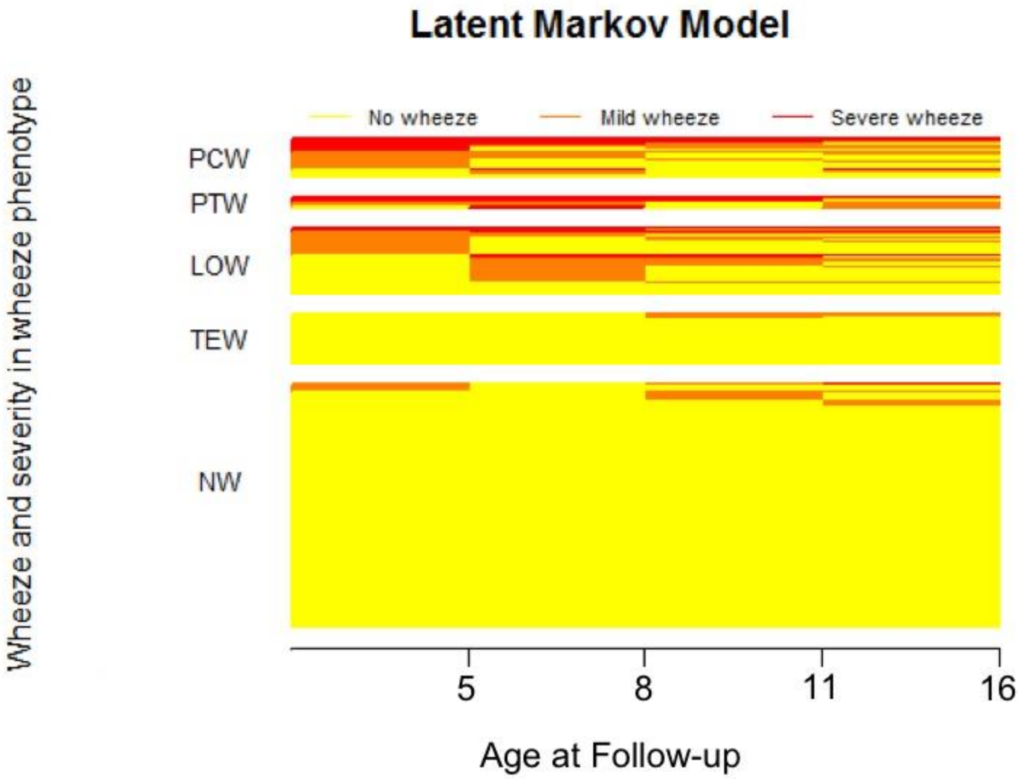


__Figure 6.3:__ Estimated posterior distribution of the latent states for each subject and each time occasion stratified by wheeze phenotype.

6.6 Discussion

6.6.1 Main findings

Our results suggest that wheeze presence and severity are dynamic processes throughout childhood. Not all children with current wheeze have a diagnosis of asthma, though this is dependent on the definition used. The use of a Latent Markov model demonstrates the influence of wheeze and symptom severity on the probability and the rate of change between latent states thereby capturing the heterogeneity of disease severity. Children with severe wheeze tend to remain in the severe wheeze state throughout childhood. Children with mild wheeze, on the other hand, tend to transition frequently throughout childhood, particularly in the no wheeze/healthy state.

Children in the severe wheeze class were significantly more likely to be sensitized and have poorer lung function, greater airway resistance, and more lung inflammation. Having a diagnosis of asthma was not significantly different in the varying wheeze states.

Children assigned to wheeze classes can exhibit varying levels of symptom severity at different times in their lives, though only a small number of children persistently wheeze throughout childhood. Severe wheeze state was present in the persistent controlled, persistent troublesome, and late onset wheeze class. A sizeable proportion of children in the persistent troublesome class, and a smaller portion in the persistent controlled class, have a low tendency of transitioning into healthier, or less severe states.

6.6.2 Limitations

The main limitation is the study is relatively small sample size. However, this sample provided the most complete dataset for us to use and strength is added with the use of data at multiple time points. This allows the LTA model to better define the transition states.[20] Further validation in other cohorts would be necessary.

We acknowledge that the definition of current wheeze is based on parental reports using standardised questionnaires. This is also true for wheeze severity, and may thus lead to overestimation.[21] However, most other epidemiological studies utilise similar definitions and we have incorporated other features of wheeze manifestations.

Finally, it is worth noting that phenotypes and 'states' discovered using data-driven methods are not observed but rather latent by nature and that the methodology is still exploratory. As such, our results should be interpreted in context and with a degree of caution.

6.6.3 Interpretation

Our study expands upon our previously published work[3] which identified four distinct wheeze classes within our population[3]. To the best of our knowledge, this is the first study that has looked at within class heterogeneity and transition probabilities using longitudinal data-driven methodology. The latent transition model is a relatively novel method of identifying how heterogeneous manifestations change over time. It differs from other longitudinal analyses of wheeze phenotypes by explicitly modelling transitions rather than trajectories, thereby avoiding the assumption that all children within a certain phenotype transition equally over time. In our model, membership of a wheeze class (phenotype) does not change as the child ages, but rather temporal changes in symptomatology becomes a characteristic of the individual phenotype and is expressed as either healthy, mild, or severe wheeze.

Our results can be compared to previous studies using similar approaches, though not all data-driven. The Melbourne Asthma study prospectively followed children with asthma from age 7 to adulthood and looked at their transitions. They found that those in the least or most severe disease groups had the highest chance of staying in those groups over time whereas those in the milder groups tended to transition between groups over time.[22] This is in concordance to the results we found and suggests that those children with severe disease have different underlying pathophysiological mechanisms that have yet to be fully understood and are likely not fully responding to current treatment strategies. These children could be considered as similar to those who have severe asthma.[23-25] Although these children make up a small proportion of the asthmatic population (10-30% depending on the cohort), they are an important subgroup as they utilise a disproportionate amount of health resources and have a significantly reduced quality of life.[24,26]

Our study can also be compared to previous studies looking at the history of wheezing patterns in early to mid-childhood years. One study, fitting separate models at each follow-up time point, found that episodic viral wheeze and multiple trigger wheeze was not stable.[27] A further study showed that children with mild disease were likely to go into remission while those with nonatopic uncontrolled wheeze transitioned to atopic uncontrolled wheeze and were less likely to go into remission.[28] Our study shows that children with severe wheeze were more likely to be atopic by late childhood. However, as previously mentioned, these studies were done by fitting separate models at each age which is in contrast to our longitudinal model. Furthermore, they did not examine transition probabilities and so within this context, our results represent an advance on previous knowledge.

Our results have similarly demonstrated that the presence of wheeze is very dynamic in each phenotype and that ascertaining severity through symptom expression shows a better picture. Furthermore, one can see from our data as well that a proportion of children in each phenotype have more severe disease than the other members suggesting that these children may actually represent different entities that do not fit the identified patterns. Despite having a high probability of belonging to a certain phenotype, children display varying levels of severity. Oksel et al[10] also found that the majority of children in all phenotypes had diminished FEV_1/FVC, though much more so in the persistent wheeze class which corresponds to a study by

Belgrave et al who identified a persistently low longitudinal FEV_1 trajectory in throughout childhood associated with recurrent wheeze and severe wheeze exacerbations.[29] This is an important aspect to be aware of as recent studies have shown that children with low FEV_1 and persistent asthma in childhood are at risk of developing COPD in adulthood.[30-32] Our results show that children in the severe wheeze state have lower FEV_1 and diminished FEV_1/FVC by age 16 compared to children in the mild/moderate wheeze state and the children that frequently transition between states suggesting that perhaps this is the group of children we should target with personalised treatment strategies in order to prevent the worsening of disease and possible subsequent development of COPD. Nevertheless, it is this pattern of wheeze heterogeneity and late outcomes that should be monitored closely in order to develop preventative strategies that can be utilised in early life. It is interesting to note that there was no significant difference in the presence of an asthma diagnosis among the different wheeze states suggesting that severity of wheeze is not a good indicator of whether someone develops asthma.

In conclusion, with the use of relatively novel data-driven methodology in this field, we have shown that presence or absence of wheeze is not a robust enough feature to ascertain phenotypes of disease. Furthermore, wheeze phenotypes are not static but rather dynamic processes involving children transitioning between states of no symptoms to mild/moderate and severe symptoms. This analysis has given us a framework of what is happening at the population level and that the pattern of severity can be identified using multiple clinical symptoms as markers. Further research is needed to identify what is occurring at an individual level in order to derive personalised approaches to prevention and treatment strategies.

6.7 References

1. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and wheezing in the first six years of life. The Group Health Medical Associates. *N Engl J Med* 1995; **332**(3): 133-8.
2. Lowe LA, Simpson A, Woodcock A, et al. Wheeze phenotypes and lung function in preschool children. *Am J Respir Crit Care Med* 2005; **171**(3): 231-7.
3. Belgrave DC, Simpson A, Semic-Jusufagic A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. *J Allergy Clin Immunol* 2013; **132**(3): 575-83 e12.
4. Henderson J, Granell R, Heron J, 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.
5. Savenije OE, Granell R, Caudri D, et al. Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA. *J Allergy Clin Immunol* 2011; **127**(6): 1505-12 e14.
6. Deliu M, Belgrave D, Sperrin M, Buchan I, Custovic A. Asthma phenotypes in childhood. *Expert Rev Clin Immunol* 2016: 1-9.
7. Oksel C, Haider S, Fontanella S, Frainay C, Custovic A. Classification of Pediatric Asthma: From Phenotype Discovery to Clinical Practice. *Front Pediatr* 2018; **6**: 258.
8. Spycher BD, Silverman M, Brooke AM, Minder CE, Kuehni CE. Distinguishing phenotypes of childhood wheeze and cough using latent class analysis. *Eur Respir J* 2008; **31**(5): 974-81.
9. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. *Eur Respir J* 2012; **40**(1): 55-60.
10. Oksel C, Granell R, Haider S, et al. Distinguishing Wheezing Phenotypes from Infancy to Adolescence: A Pooled Analysis of Five Birth Cohorts. *Ann Am Thorac Soc* 2019.
11. Reddel HK, Bateman ED, Becker A, et al. A summary of the new GINA strategy: a roadmap to asthma control. *Eur Respir J* 2015; **46**(3): 622-39.
12. Network BTSSIG. British guideline on the management of asthma: A national clinical guideline. *Thorax* 2014; (69): 1-192.
13. Soto-Ramirez N, Ziyab AH, Karmaus W, et al. Epidemiologic methods of assessing asthma and wheezing episodes in longitudinal studies: measures of change and stability. *J Epidemiol* 2013; **23**(6): 399-410.
14. Garden FL, Simpson JM, Mellis CM, Marks GB, Investigators C. Change in the manifestations of asthma and asthma-related traits in childhood: a latent transition analysis. *Eur Respir J* 2016; **47**(2): 499-509.
15. Lowe L, Murray CS, Custovic A, et al. Specific airway resistance in 3-year-old children: a prospective cohort study. *Lancet* 2002; **359**(9321): 1904-8.
16. Custovic A, Simpson BM, Murray CS, et al. The National Asthma Campaign Manchester Asthma and Allergy Study. *Pediatr Allergy Immunol* 2002; **13 Suppl 15**: 32-7.
17. Bartolucci F, Farcomeni A, Pennoni F. Latent Markov Models for Longitudinal Data: Chapman & Hall/CRC, Boca Raton; 2013.
18. Bartolucci F, Pandolfi S, Pennoni F. LMest: An R Package for Latent Markov Models for Longitudinal Categorical Data. *Journal of Statistical Software; Vol 1, Issue 4 (2017)* 2017.

19. Woodcock A, Lowe LA, Murray CS, et al. Early life environmental control: effect on symptoms, sensitization, and lung function at age 3 years. *Am J Respir Crit Care Med* 2004; **170**(4): 433-9.

20. Velicer WF, Martin RA, Collins LM. Latent transition analysis for longitudinal data. *Addiction* 1996; **91 Suppl**: S197-209.

21. Lowe L, Murray CS, Martin L, et al. Reported versus confirmed wheeze and lung function in early life. *Arch Dis Child* 2004; **89**(6): 540-3.

22. Wolfe R, Carlin JB, Oswald H, Olinsky A, Phelan PD, Robertson CF. Association between allergy and asthma from childhood to middle adulthood in an Australian cohort study. *Am J Respir Crit Care Med* 2000; **162**(6): 2177-81.

23. Bossley CJ, Fleming L, Ullmann N, et al. Assessment of corticosteroid response in pediatric patients with severe asthma by using a multidomain approach. *J Allergy Clin Immunol* 2016; **138**(2): 413-20 e6.

24. Fitzpatrick AM. Severe Asthma in Children: Lessons Learned and Future Directions. *J Allergy Clin Immunol Pract* 2016; **4**(1): 11-9; quiz 20-1.

25. Chung KF. New treatments for severe treatment-resistant asthma: targeting the right patient. *Lancet Respir Med* 2013; **1**(8): 639-52.

26. Chastek B, Korrer S, Nagar SP, et al. Economic Burden of Illness Among Patients with Severe Asthma in a Managed Care Setting. *J Manag Care Spec Pharm* 2016; **22**(7): 848-61.

27. Schultz A, Devadason SG, Savenije OE, Sly PD, Le Souef PN, Brand PL. The transient value of classifying preschool wheeze into episodic viral wheeze and multiple trigger wheeze. *Acta Paediatr* 2010; **99**(1): 56-60.

28. Just J, Saint-Pierre P, Gouvis-Echraghi R, et al. Wheeze phenotypes in young children have different courses during the preschool period. *Ann Allergy Asthma Immunol* 2013; **111**(4): 256-61 e1.

29. Belgrave DCM, Granell R, Turner SW, et al. Lung function trajectories from pre-school age to adulthood and their associations with early life factors: a retrospective analysis of three population-based birth cohort studies. *Lancet Respir Med* 2018; **6**(7): 526-34.

30. Grad R, Morgan WJ. Long-term outcomes of early-onset wheeze and asthma. *J Allergy Clin Immunol* 2012; **130**(2): 299-307.

31. Bui DS, Lodge CJ, Burgess JA, et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. *Lancet Respir Med* 2018; **6**(7): 535-44.

32. Berry CE, Billheimer D, Jenkins IC, et al. A Distinct Low Lung Function Trajectory from Childhood to the Fourth Decade of Life. *Am J Respir Crit Care Med* 2016; **194**(5): 607-12.

6.8 Supplementary Material

6.8.1 Methods

Study design: Unselected birth cohort

Setting: A mixed urban-rural population within 50 square miles of South Manchester and Cheshire, United Kingdom located within the maternity catchment area of Wythenshawe and Stepping Hill Hospitals

Screening and recruitment: All pregnant women were screened for eligibility at antenatal visits (8-10th week of pregnancy). Of the 1499 couples who met the inclusion criteria (≤10 weeks of pregnancy, maternal age ≥18 years, and questionnaire and skin prick data test available for both parents), 288 declined to take part in the study and 27 were lost to follow-up between recruitment and the birth of a child. A total of 1184 children born into the study had at least some evaluable data.

Definition of variables

Current rhinitis: Positive answer to "In the past 12 months, has your child had a problem with sneezing or a runny or blocked nose when he/she did not have a cold or the flu?"

Current eczema: Positive answer to "Has your child had eczema within the past 12 months?"

Atopic sensitisation: Mean diameter (MWD) 3mm larger than the negative control to at least one allergen.

Maternal /Paternal Smoking: Positive answer to "Does your child/'s mother, father smoke?"

Lung function, airway hyperreactivity and airway inflammation

We measured lung function using spirometry at ages 8, 11 and 16 years using a Lilly pneumotachograph system with animated incentive software (Jaeger, Würzburg, Germany), or for home visits, a flow turbine spirometer (Micro Medical, UK)[1]. FEV_1 % predicted[2] and FEV_1/FVC ratio were recorded. Specific airway resistance (sRaw) was measured using whole-body plethysmography (Masterscreen Body 4.34; Jaeger, Würzburg, Germany) plethysmography at ages 3, 5, 8, 11, 16.[3,4] Airway hyperreactivity (AHR) was measured using standard quadrupling doses of methacholine in a 5-stage process at ages 8 and 11 years.[5] Children were considered to have AHR if there was a 20% decrease in FEV_1 by the final stage (16mg/mL). We also calculated a dose-response slope.[6] Airway inflammation was recorded at age 8, 11, and 16years as a measure of Fractional Exhaled nitric oxide (FeNO) and performed according to the American Thoracic Society guidelines using either a chemiluminescence analyser or an electrochemical analyser (NIOX, Solna, Sweden) .[7] Data were expressed in parts per billion (ppb).

6.8.2 Results

Table E6.1: *Descriptive characteristics of the remaining 506 children without a complete dataset*

Boys (whole population) Girls (whole population)	357, 71% 282, 39%				
Family history of asthma	134, 26%				
Family history of atopy	392, 77%				
	Age 3	**Age 5**	**Age 8**	**Age 11**	**Age 16**
BMI, median (IQR)	16.1 (15.2-17.9)	16.3 (15-16.9)	16.6 (15.8-18.1)	18.5 (16.9-20.5)	21.4 (19.8-23.7)
Maternal smoking	125, 25%	125, 25%	97, 19%	58, 11%	20, 4%
Paternal smoking	158, 31%	155, 31%	118, 23%	77, 15%	18, 4%
Asthma ever	92, 18%	159, 31%	100, 20%	98, 19%	63, 12%
Atopic sensitization (SPT)	114, 23%	156, 31%	156, 31%	116, 23%	45, 9%
Current rhinitis	n/a	147, 29%	136, 27%	136, 27%	52, 10%
Current eczema	131, 26%	165, 33%	120, 24%	70, 14%	22, 4%
Any current asthma medication	84, 17%	165, 33%	103, 20%	95, 19%	36, 7%
Current wheeze, parentally reported	153, 30%	146, 29%	94, 19%	90, 18%	30, 6%
Current asthma, parentally reported			63, 12%	64, 12%	24, 5%
	Asthma severity symptoms				
Number of wheeze attacks					
none	n/a	280, 55%	281, 56%	212, 42%	476, 94%
1-3	n/a	68, 13%	48, 9%	34, 7%	15, 3%
4 or more	n/a	49, 10%	23, 4%	30, 6%	13, 3%
Exercise induced wheeze		58, 11%	35, 6%	42, 8%	22, 4%
Wheeze affecting sleep					
none	n/a	317, 63%	306, 60%	238, 47%	495, 98%
<1 night per week	n/a	33, 7%	31, 6%	24, 5%	9, 2%
More than 1 night per week	n/a	34, 7%	14, 3%	16, 3%	2, 1%
Wheeze limiting speech	n/a	31, 6%	8, 2%	61, 11%	0
Children considered to be severe	n/a	28, 6%	7, 1%	14, 3%	0

Latent Markov Models to investigate patterns of wheeze

We firstly applied a univariate LM model to investigate the prevalence of wheeze over time where we look at the transition analysis and its accompanying measurement errors. The main assumption is that the "true state" characterizing a sample unit at a certain occasion may be affected by measurement error. The "true states" corresponding to the latent states and transitions between each pair of these states are studied through the transition probabilities[8]. In this study, we interpret the latent states as the true presence or absence of wheeze. By looking at the transition probabilities, we can verify how each subject progresses in the true wheeze state, whereas the conditional response probabilities indicate how the reported current wheeze state depends on the true one.

Model selection was based on the BIC index (Table E6.5), which indicated a 2-state solution with homogeneous transition probabilities.

Table E6.6 shows the estimates of the conditional response probabilities, whereas the estimates of the initial and transition probabilities are reported in Tables E6.7 and E6.8, respectively. Results in Table E6.6 show that the latent states are related to the presence of wheeze. The first state clearly corresponds to the absence of wheeze symptoms, while the second state corresponds to an increased tendency of wheezing. Results in Table E6.7 show that the majority of subjects belong to the first latent state, no wheeze, at the beginning of the study. Transition probabilities (Table E6.8) show high persistence in state 1 (no wheeze) and moderate persistence in state 2 (current wheeze).

Table E6.2: *Multivariate latent Markov model selection using BIC index*

	Number of states	BIC
Time homogeneous	1	7197.96
	2	4241.27
	3	4187.42
Time heterogeneous	1	7197.96
	2	4258.83
	3	4211.29

Table E6.3: *Estimates of the initial probabilities*

States	Initial probabilities
1	0.129
2	0.055
3	0.816

Table E6.4: *Estimates of the transition probabilities between the latent states.*

States	1	2	3
1	0.360	0.032	0.608
2	0.125	0.655	0.220

3	0.068	0.013	0.919

Table E6.5: *Model selection using BIC index.*

	Number of states	BIC
Time homogeneous	1	3093.45
	2	2617.95
Time heterogeneous	1	3093.46
	2	2645.57

Table E6.6*: Estimates of the conditional response probabilities.*

	Current wheeze	
States	**0**	**1**
1	0.966	0.034
2	0.266	0.735

Table E6.7: *Estimates of the initial probabilities.*

States	Initial probabilities
1	0.714

2	0.286

Table E6.8: *Estimates of the transition probabilities between the latent states.*

States	1	2
1	0.969	0.031
2	0.223	0.777

6.8.3 References

1. Beydon N, Davis SD, Lombardi E, et al. An official American Thoracic Society/European Respiratory Society statement: pulmonary function testing in preschool children. Am J Respir Crit Care Med 2007;175:1304-45.
2. Stanojevic S, Wade A, Cole TJ, et al. Spirometry centile charts for young Caucasian children: the Asthma UK Collaborative Initiative. Am J Respir Crit Care Med 2009;180:547-52.
3. Custovic A, Simpson BM, Murray CS, et al. The National Asthma Campaign Manchester Asthma and Allergy Study. Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology 2002;13 Suppl 15:32-7.
4. Lowe LA, Woodcock A, Murray CS, Morris J, Simpson A, Custovic A. Lung function at age 3 years: effect of pet ownership and exposure to indoor allergens. Arch Pediatr Adolesc Med 2004;158:996-1001.
5. Crapo RO, Casaburi R, Coates AL, et al. Guidelines for methacholine and exercise challenge testing-1999. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. Am J Respir Crit Care Med 2000;161:309-29.
6. Beydon N, Pin I, Matran R, et al. Pulmonary function tests in preschool children with asthma. Am J Respir Crit Care Med 2003;168:640-4.
7. 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;171:912-30.
8. Bartolucci F, Farcomeni A, Pennoni F. Latent Markov Models for Longitudinal Data: Chapman & Hall/CRC, Boca Raton; 2013.

Part 3: Concluding Remarks

Chapter 7 Discussion

Chapter specific discussions and conclusions have been given at the end of each chapter and will not be expanded upon here. Rather, this discussion chapter aims to relate the findings back to the research questions and provide some ideas for future work.

7.1 Research question 1: Can we use data-driven methods to uncover patterns among asthma datasets and how can this help guide our further understanding of the disease?

Through this PhD, we have demonstrated that the use of data-driven machine learning methods such as cluster analysis, latent class analysis and latent transition analysis can be useful tools in ascertaining subtypes of disease. Chapter 1 of this PhD brings this all into context by describing the evolution of ascertaining asthma phenotypes and subtypes; first using expert observations and then evolving to the use of data driven methods. The various work described has helped identify clinically meaningful asthma subtypes. The number and types of clusters can vary depending on the sample size, timing of follow-up, data transformations, or method which has largely contributed to somewhat different study conclusions. However, this has also revealed hidden structures within complex data sets that contribute to the heterogeneity between subjects within specific classes (or subtypes) that should, in theory, be homogenous.

Chapter 2 introduced the current knowledge on asthma phenotypes in childhood and the implications of clinical subtyping of childhood asthma. Briefly, various approaches, from modelling single or multiple symptoms (such as wheeze), lung function, biomarkers, allergic sensitization, genetics, to multi-variate models cross-sectionally or longitudinally have identified different subtypes of asthma. The information gathered and presented in this chapter demonstrates that there is a clear lack of standardization of statistical methods and which variables to use for defining childhood wheezing and asthma, which has led to inconsistencies in findings relating to genetic and environmental risk factors.

7.2 Research question 2: What main features of the asthma syndrome can be used to ascertain the heterogeneity of the disease?

Within the scope of this PhD, using an almost complete dataset from a population cohort and applying a framework of blending data driven methods and clinical expertise, we identified four main features that are useful in disaggregating asthma subtypes: age of onset, atopy, exacerbations, and severity; the results of which are described in chapter 4. Our conclusion from this set of analyses is that not every variable in a dataset is informative when ascertaining possible asthma subtypes. Our hierarchical cluster model did not yield stable clusters when inputting all variables despite a general acceptance that models become more stable with linear increases in variables. When performing dimensionality reduction in order to reduce all variables to a smaller number of uncorrelated components that should, in theory, retain as much information about the dataset as possible, our resulting hierarchical cluster model was still unstable. Instead, our semi-automated approach where clinical experts compared the post-analysis results from the two HC models identified four informative features/domains of the disease that markedly increased cluster stability and produced a more clinically meaningful picture. Putting this into context in order to disaggregate the heterogeneity of asthma, common questions used to determine the presence of disease in other studies may not be informative when uncovering asthma subtypes and so in chapter 4 we propose a framework focusing on the domains we identified.

Within the scope of these four domains, our study identified an exacerbation-prone cluster that is likely a separate endotype with a unique, still not fully understood, aetiology that is not solely characterised by asthma severity or symptom control, as even decreasing symptoms may not always mean a decrease in the number of exacerbations.[1,2] This has significant potential in targeting exacerbation prevention strategies particularly as a recent clinical trial could not demonstrate the effect of increasing the dose oh inhaled corticosteroids when symptom control worsens prior to an attack in order to prevent the impending exacerbation.[3]

Severity was another one of the key features found to be informative for asthma subtyping. Among the clusters, severity was pretty evenly split with children with the most severe asthma found in each cluster. This finding, in line with other recent studies[4,5] showing even distribution of severity among clusters rather than an independent cluster.

Despite a common acceptance among the clinical community that lung function is a key determinant of asthma, our analysis actually found that the majority of children had normal lung function suggesting that lung function is likely less important for subtyping.

This cross-sectional study gave us a snapshot of the heterogeneity that was then expanded on in subsequent chapters.

This is evidenced in the next two chapters (5 and 6) where we looked at some of the key features (severity and exacerbations) in greater detail using a well-defined, comprehensive, population-based birth cohort.

7.3 Research question 3: How can we exploit the wealth of data provided by longitudinal birth cohorts in order to understand the severity of asthma?

The use of birth cohort studies has provided the additional benefit of allowing us to incorporate systematic observation prior to disease onset which facilitates the exploration of the natural history of disease. With longitudinal, repeated-measure data from birth cohorts, the development of disease can be followed over time, which essentially parallels clinical diagnosis and follow-up observations. Additionally, it allows researchers to estimate distributions and prevalence rates in the population as well as attributable risks. A representative sample of the population is the ideal setting to analyse relationships between subjects and confounders to exposures and outcomes.[6]

Longitudinal studies allow subjects to be followed prospectively over time in order to monitor risk factors and/or observed outcomes. It allows us to see the development or changes of certain population characteristics over time. Chapters 5 and 6 are extensions from chapter 4 as the goal was to try and map the trajectories of each of these domains in a well-defined birth cohort. Chapter 5 of this PhD focused on wheeze exacerbations and utilised longitudinal data from primary care data associated with four follow-up points to identify two distinct independent trajectories of wheeze exacerbations: infrequent and frequent exacerbations. Children who frequently exacerbate are more likely to require more inhaled therapy for symptom control, be sensitized, and have a diagnosis of asthma by late childhood. These children are also more likely to have poorer lung function, though it is important to look not only at FEV_1 (as the majority of our children had normal FEV_1 values) but rather the ratio of FEV_1/FVC which would be a better predictor of those at risk of exacerbations. This is a crucial finding as it has been shown that children with recurrent early life exacerbations were more likely to persist in having poor lung function by adulthood which is also an important risk factor in the development of chronic obstructive pulmonary disease (COPD).[7] Duration of breastfeeding (those with minimal breastfeeding) was found to be the strongest early life risk factor for exacerbations. Given that many studies have shown the protective nature of breastfeeding on asthma development, asthma severity, and asthma exacerbations[8-10], this modifiable aspect could be prevented in order to decrease the associated risk.

However, another key finding is that a large proportion of children with well-controlled mild asthma and normal lung function still have frequent exacerbations implying that there are likely some unknown underlying pathophysiological mechanisms that may contribute to this. It is thus possible that exacerbation prone asthma may represent an independent subtype. Although the sample size was small, limiting the interpretation of results, the proportion of exacerbating children is a representative sample of our population in line with other reported studies. Furthermore, strength is attributed to the fact that data was extracted from primary care records and not solely from parentally recollected answers to questionnaires.

Using a relatively novel method in this field, latent transition analysis based on a latent Hidden Markov model, we demonstrate in chapter 6 that wheeze severity is a dynamic process throughout childhood and that the majority of children transition between different states of severity. The latent transition model is able to identify how heterogeneous manifestations change by modelling transitions over time. This chapter showed how temporal changes in wheeze symptoms are actually characteristics of individual wheeze phenotypes in the form of healthy, mild, or severe wheeze. Children who start off in the severe state early on tend to remain in the severe state by late childhood. These children are also more likely to have poorer lung function, more airway inflammation and resistance. This suggests that different mechanisms are responsible in contributing to a variable response to current treatment strategies with inhaled corticosteroids.[11,12] It has been suggested that severe asthma in childhood is characterised by steroid-insensitive eosinophilia likely owing to the lack of successful response.[13,14] Children with mild wheeze tend to transition frequently particularly within no wheeze/no symptom state. This suggests that the current methods to ascertain wheeze patterns (using presence or absence of wheeze) are not robust enough to capture the heterogeneity of wheeze. When looking at wheeze phenotype classes from previously published work within MAAS[15], this model has shown that wheeze classes ascertained through standard latent class analysis are not actually static and that children with persistent troublesome wheeze have similar severity patterns as those with persistent controlled wheeze. Furthermore, despite common assumptions, wheeze severity is not a good indicator of whether or not a child will have a diagnosis of asthma.

Age of onset of wheeze, as one of the other key features, has already been mentioned in previous chapters and was derived by Belgrave et al[15] using latent class analysis and so this was not expanded on in the thesis. Briefly, they discovered four main wheeze phenotypes: transient early, late onset, persistent controlled, persistent troublesome. Results from chapter 4 have shown that age of wheeze onset is a key discriminator for endotypes however children assigned to each class show different levels of symptom severity as chapter 6 has demonstrated. Although the majority of severe wheeze was in the persistent wheeze classes (controlled and troublesome), only a small proportion of children actually remained in the severe wheeze state during childhood and the greatest amount of severity occurred early on. This can, in one way, be explained by the preschool effect whereby children are first exposed to infections and viruses which lead to the development of respiratory symptoms.[16] Within the other classes, the majority of children either mild or no symptoms. This suggests that labelling children as having specific phenotypes is not the way forward as there is a vast amount of within class heterogeneity. Children can transition through phenotypes highlighting the importance of age when defining asthma subgroups. Results from this thesis suggest that we should not solely use the presence or absence of wheeze as a marker of wheeze trajectory as it is evident that a large proportion of children do not fit their allocated wheeze phenotype patterns. Instead we should be looking at different symptom expressions and how it varies from age to age.

With respect to children who frequently exacerbate, not surprisingly, we found that they were more likely to have persistent wheeze suggesting that there are likely undiscovered pathophysiological mechanisms working synergistically. One explanation could be that there are multiple triggers contributing. A study by Spycher et al showed that multi trigger wheeze (MTW) was associated with higher severity wheeze (which includes those children who have frequent attacks) and that it tends to persist throughout childhood.[17]

This thesis did not look at the atopy domain separately as it has been previously investigated. Using a Hidden Markov Model (a model that our LTA in Chapter 6 is also based on), Simpson et al from the MAAS group identified four atopy phenotypes: multiple early, multiple late, dust mite, and non-dust mite, along with a no latent variability class. Multiple early sensitization phenotype was highly associated with asthma, significantly reduced lung function, and increased risk of hospital admissions.[18] Although these clusters were not incorporated into an analysis within the thesis as the numbers were too small, we used the presence or absence of any sensitization variable as a late childhood outcome. We found that children with severe wheeze state and children with frequent exacerbations were more likely to be sensitised by late childhood. This is an important finding as a recent study showed that children with atopic uncontrolled wheeze were more likely to continue having high symptom burden.[19] This also reflects the acquisition of allergic sensitization with age.[20] Furthermore, children with frequent exacerbations were more likely to be sensitized suggesting that similar risk factors and mechanisms may be contributing to atopy and exacerbations. This has implications on treatment strategies as it has been found that low interferon induction and high proinflammatory cytokine response to rhinovirus-16 increased the risk of wheeze and lower respiratory tract infections in early life[21], which has been found to drive frequent asthma exacerbation phenotypes.[22,23] Children who display this lack of type I/II interferon immune response are also more likely to be sensitized throughout childhood due to a synergistic weak Th2 cytokine response to PHA thereby increasing susceptibility to virus infection.[21] Further research is required in identifying the underlying immunophenotype in order to develop novel biologic therapies.

It is evident that the same pattern of symptoms among different children does not necessarily equate to the same underlying pathophysiological mechanisms underpinning this heterogeneity. Given that different mechanisms are not exclusively independent, it is entirely possible that children with the most severe disease expression may indeed have multiple mechanisms working synergistically.

There is a great amount of diversification in results among different studies owing to different number of clusters, sample sizes, frequency and timing of data collection, and statistical methodology which have demonstrated variability in study conclusions. Results from this thesis indicates considerable heterogeneity among children within the same classes which could be used as an explanation for the lack of consistent results. It is also likely that different pathophysiological mechanisms are responsible for expressing similar patterns of symptoms among different children.

Placing patients in a ‘one size fits all’ treatment box is no longer becoming optimal as it has demonstrated not only a large proportion of treatment failure but an increase in the risk side effect vs benefit ratio. We need to move away from symptom/diagnosis/lung function based treatments towards mechanism-based strategies. Data driven methods have been paramount in identifying patterns within datasets that cannot be seen by simple observation. This thesis has shown how machine learning can facilitate endotype discovery by identifying latent (hidden) structures within the chosen datasets. However, there is a lack of certainty in calling these subgroups (clusters/classes) ‘true endotypes’ and so results need to be interpreted in context as they are not yet translatable to clinical care. To put in perspective, a randomised control study could not show significantly different treatment responses when children were grouped into clusters suggesting that treatment is likely not endotype-specific.[24] What is needed is pooling of data not only from observational birth cohorts, but also from different patient studies and randomised control trials in order to triangulate the knowledge obtained to pinpoint disease mechanisms, biomarkers, genetics, predictors of future disease, and response to treatment.[25] The emergence of biologic treatments can substantially improve our endotype discovery by identifying treatment responders and therefore detecting possible underlying pathways of disease expression in different subgroups.[26] Therefore integration of clinical trial and cohort data, endotypes, biomarkers, -omics data, *in vitro* mechanistic studies using human samples, in vivo experimental models, along with advanced computer algorithms using 'big data', could help identify novel targets and bring about true stratified personalised medicine. This would bring us one step closer from analysing data at a population level to analysing at a person level.

7.4 Future work

While each chapter of this thesis presents important advances in understanding the asthma syndrome, several areas warrant further investigation. Future work that is beyond the scope of this PhD will consist of the following:

1) Creating a framework and model that could combine the four features (age of onset, atopy, exacerbations, severity). Given the complexity of the different variables (mixed data consisting of ordinal, binary, continuous and categorical variables), there is yet to be a model that can handle this as it would have to encompass all original variables used in ascertaining the clusters and classes. It is not yet possible to combine the different classes/clusters as variables since they were identified using cluster analysis and latent class analysis (LCA) and using different clinical follow-up points. As mentioned previously, LCA allocates probabilities of group membership while cluster analysis divides based on distance measures and so conceptually, this is very difficult to comprehend. However, it would be very informative to see how each child moves from one class to another.
2) Replicating these results in other cohorts, particularly within STELAR (Study Team for Early Life Asthma Research)[27] consortium. However, data from the other cohorts has been collected at different timepoints, some measures are completely missing in some studies, different drop-out rates and at different time points among studies, and the

lack of precise information on symptoms and timing of exacerbations. The challenge lies in finding algorithms that could blend all this data.

3) Blending in results from different study types (birth cohorts, patient studies, randomised control trials) with genomics, transcriptomics, and proteomics in order to further disaggregate the scope of the asthma syndrome. The use of biomarkers could enable targeted prescribing methods.

7.5 Conclusion

In conclusion, this thesis has shown that data driven algorithms such as cluster analysis, K-means longitudinal analysis, and latent transition analysis are useful methods in disaggregating the heterogeneity of the asthma syndrome. Utilising these methods in clinical context through different rich data sources, we have identified main features useful in identifying subtypes of disease. We identified exacerbation prone asthma in one cohort and mapped the trajectory throughout childhood through two independent subgroups in a birth cohort. The relatively novel method used to identify wheeze severity patterns could be used as a future methodological framework for visualising the dynamic processes of respiratory symptoms throughout childhood. Overall, this thesis has demonstrated the utility of data driven methods in understanding the patterns of asthma symptoms in childhood.

7.6 References

1. Kupczyk M, ten Brinke A, Sterk PJ, et al. Frequent exacerbators--a distinct phenotype of severe asthma. Clin Exp Allergy 2014;44:212-21.

2. O'Byrne PM, Barnes PJ, Rodriguez-Roisin R, et al. Low dose inhaled budesonide and formoterol in mild persistent asthma: the OPTIMA randomized trial. Am J Respir Crit Care Med 2001;164:1392-7.

3. Jackson DJ, Bacharier LB, Mauger DT, et al. Quintupling Inhaled Glucocorticoids to Prevent Childhood Asthma Exacerbations. N Engl J Med 2018;378:891-901.

4. Fitzpatrick AM, Teague WG, Meyers DA, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. J Allergy Clin Immunol 2011;127:382-9 e1-13.

5. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. Eur Respir J 2012;40:55-60.

6. Szklo M. Population-based cohort studies. Epidemiol Rev 1998;20:81-90.

7. Belgrave DCM, Granell R, Turner SW, et al. Lung function trajectories from pre-school age to adulthood and their associations with early life factors: a retrospective analysis of three population-based birth cohort studies. Lancet Respir Med 2018;6:526-34.

8. Azad MB, Vehling L, Lu Z, et al. Breastfeeding, maternal asthma and wheezing in the first year of life: a longitudinal birth cohort study. Eur Respir J 2017;49.

9. Ahmadizar F, Vijverberg SJH, Arets HGM, et al. Breastfeeding is associated with a decreased risk of childhood asthma exacerbations later in life. Pediatr Allergy Immunol 2017;28:649-54.

10. Dogaru CM, Nyffenegger D, Pescatore AM, Spycher BD, Kuehni CE. Breastfeeding and childhood asthma: systematic review and meta-analysis. Am J Epidemiol 2014;179:1153-67.

11. Szefler SJ, Phillips BR, Martinez FD, et al. Characterization of within-subject responses to fluticasone and montelukast in childhood asthma. J Allergy Clin Immunol 2005;115:233-42.

12. Lemanske RF, Jr., Mauger DT, Sorkness CA, et al. Step-up therapy for children with uncontrolled asthma receiving inhaled corticosteroids. N Engl J Med 2010;362:975-85.

13. Miranda C, Busacker A, Balzar S, Trudeau J, Wenzel SE. Distinguishing severe asthma phenotypes: role of age at onset and eosinophilic inflammation. J Allergy Clin Immunol 2004;113:101-8.

14. Phipatanakul W, Mauger DT, Sorkness RL, et al. Effects of Age and Disease Severity on Systemic Corticosteroid Responses in Asthma. Am J Respir Crit Care Med 2017;195:1439-48.

15. Belgrave DCM, Simpson A, Semic-Jusufagic A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. J Allergy Clin Immunol 2013;132:575-83 e12.

16. Ball TM, Castro-Rodriguez JA, Griffith KA, Holberg CJ, Martinez FD, Wright AL. Siblings, day-care attendance, and the risk of asthma and wheezing during childhood. N Engl J Med 2000;343:538-43.

17. Spycher BD, Cochrane C, Granell R, et al. Temporal stability of multitrigger and episodic viral wheeze in early childhood. Eur Respir J 2017;50.

18. Simpson A, Tan VY, Winn J, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. Am J Respir Crit Care Med 2010;181:1200-6.
19. Just J, Saint-Pierre P, Gouvis-Echraghi R, et al. Wheeze phenotypes in young children have different courses during the preschool period. Ann Allergy Asthma Immunol 2013;111:256-61 e1.
20. Kulig M, Bergmann R, Klettke U, Wahn V, Tacke U, Wahn U. Natural course of sensitization to food and inhalant allergens during the first 6 years of life. J Allergy Clin Immunol 1999;103:1173-9.
21. Custovic A, Belgrave D, Lin L, et al. Cytokine Responses to Rhinovirus and Development of Asthma, Allergic Sensitization, and Respiratory Infections during Childhood. Am J Respir Crit Care Med 2018;197:1265-74.
22. Gill MA, Bajwa G, George TA, et al. Counterregulation between the FcepsilonRI pathway and antiviral responses in human plasmacytoid dendritic cells. J Immunol 2010;184:5999-6006.
23. Wark PA, Johnston SL, Bucchieri F, et al. Asthmatic bronchial epithelial cells have a deficient innate immune response to infection with rhinovirus. J Exp Med 2005;201:937-47.
24. Chang TS, Lemanske RF, Jr., Mauger DT, et al. Childhood asthma clusters and response to therapy in clinical trials. J Allergy Clin Immunol 2014;133:363-9.
25. Saglani S, Custovic A. Childhood Asthma: Advances Using Machine Learning and Mechanistic Studies. Am J Respir Crit Care Med 2019;199:414-22.
26. Custovic A, Henderson J, Simpson A. Does understanding endotypes translate to better asthma management options for all? J Allergy Clin Immunol 2019;144:25-33.
27. Custovic A, Ainsworth J, Arshad H, et al. The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together. Thorax 2015;70:799-801.

7.7 Addendum: The Academic Journey

7.7.1 Evolution of the statistical journey

The rapid review of chapter 2 led to the realisation that there were 2 main aspects surrounding the use of machine learning data-driven methods in asthma studies: pre-processing of the data and the choice of cluster method. Healthcare data is comprised of mixed type of data and it has been shown in this PhD that different pre-processing methods can lead to heterogeneous results. The most commonly used methods of choosing the variables for the model are generally subjective investigator choice and dimensionality reduction using principal component analysis or factor analysis (depending on the dataset used). The next step is to choose the clustering method, and the most commonly used methods are hierarchical clustering using Ward's method, k-means clustering, and latent class analysis. Only k-means and latent class analysis can be also used longitudinally. Hierarchical clustering and k-means (this method is very sensitive to outliers and numerical variables) assign subjects to only one cluster, whereas latent class analysis assigns probabilities of class allocation which can display the uncertainties as some subjects may be borderline between classes. However, all of these methods assume that the resulting subgroups are static.

For both data selection and clustering method, there is no standardised and validated choice for 'best practice' and so chapter 4 was based on the most commonly used methods which allowed me to understand my dataset best. The first approach was to reduce the dataset into smaller, informative components as, in theory, this should provide the most stable results. However, the resulting clusters were not stable and so the model was re-run with all variables in raw format, yet this yielded similar results. A decision was made to then look at the resulting clusters from both models simultaneously with four experts in order to reduce inter-researcher variability and bias. It became evident that four main features were driving the cluster allocation which were: age of onset, atopy, asthma exacerbations, and asthma severity. In order to look at each of these domains in more detail, a birth cohort would provide the better platform as chapter 4 was a cross sectional analysis only.

Chapter 5 aimed at identifying trajectories of exacerbations throughout childhood. The methodological choice for this was a challenging one as I did not want to categorise exacerbations (i.e. categorising to ≥3 assumes that a child with 3 exacerbations is the same as a child with 10, which is not necessarily true), but rather keep them in their raw numerical format. The most applicable cluster method was k-means longitudinal (kML) which is capable of handling numeric data and the element of longitudinal time variability. The model identified two distinct exacerbation subgroups with different late childhood outcomes.

Chapter 6 aimed at identifying patterns of wheeze severity. Wheeze severity was initially defined using the British Thoracic Guidelines step in treatment as this is the most common definition found in literature. The initial methodology used was latent class analysis as it is very sensitive to categorical variables such as the step of treatment. The model identified three classes as “No wheeze” (no treatment), “mild wheeze” (BTS step 1), “moderate-severe wheeze” (BTS step 2-3). There were no children on step 4. Looking at each class closely, there were no distinguishing features and, in some cases, the lung function for children with no treatment was worse than those on treatment which was not clinically intuitive. As a result, it became evident that the use of treatment step as a marker of severity was not a robust enough way to distinguish severity subgroups in our dataset.

The use of four severe symptoms as a proxy of wheeze severity was then an evolution as symptoms are a sign of disease control and expression. Furthermore, previous research studies have made me realise that wheeze is not a static state but rather a dynamic process that likely changes over time. However, none of the clustering models that were previously used in this PhD would capture this phenomenon; a latent transition model would be most applicable. This model looks at intra group variability over time and shows how subjects transition from different states. The model identified three states of symptom expression which were labelled as “no wheeze”, “mild/moderate wheeze”, and “severe wheeze” and showed that children with mild/moderate wheeze most frequently transition through states of severity in childhood. This model was then applied to the previous wheeze phenotypes identified using latent class analysis and showed the dynamic transitions of symptom expression between each phenotype class suggesting that wheeze classes are not static.

7.7.2 Influence of external factors such as adherence and external environment on results presented

Adherence

Adherence to medication is an important aspect to consider in those children with difficult to treat severe asthma. Non-adherence is a major reason for increased asthma-related emergency admissions, poor control, increased oral steroid use, and persistent eosinophilic inflammation.[1,2] Adequate assessment of adherence would potentially differentiate genuine severe disease, inadequacy of treatment, or failure to take medication. As yet, no method exists that can monitor adherence with 100% certainty.

Adherence to inhaled corticosteroids is generally relatively poor, likely due to the absence of immediate relief.[3] Current methods of measuring adherence include prescription data, parental reports, diaries, and canister weighing, however, these are generally considered inaccurate.[4] Fraction of exhaled Nitric Oxide (FeNO), a proposed biomarker for eosinophilia/Th2 inflammation, has been gaining momentum in its possible use as a marker of both response to therapy and adherence.[5-7] Studies have shown that FeNO levels fall with directly observed inhaled corticosteroid treatment. However, it should be noted that this will apply only to those children with Th2 driven eosinophilic inflammation and above the age of 5 years due to the technique involved.

In the MAAS cohort used in this thesis, non-adherence could have accounted for some of the severity/exacerbation subgroups identified. In chapter 5, it could be argued that children with frequent exacerbations were likely not taking their inhaled corticosteroids. The greatest exacerbation burden occurred between birth and age 4 when FeNO was not able to be measured. Almost all of the children with frequent exacerbations were on inhaled corticosteroids. Although these children were more likely to have higher FeNO levels at age 8, it is not possible to conclude that this would be the same earlier in childhood. Furthermore, given that it is a birth cohort with follow-ups at different age points, it is very difficult to correctly ascertain adherence between visits without relying on parental recall and prescription data which, as stated earlier, is inaccurate. Nevertheless, it is entirely possible that the children were not using their inhalers and were thus having exacerbations, but likewise it is possible that some underlying mechanism that we are not aware of is also contributing to this effect. Consequently, it is important to be aware of the high rates of non-adherence and so a way forward would be to design a study incorporating strict adherence criteria (i.e. using electronic monitoring devices +/- FeNO).

External factors

This thesis has shown that there is a lot of heterogeneity in asthma, however, data on social and environmental factors was not available. Several studies have shown that socioeconomic status (SES) may be a risk factor for the development of asthma.[8-13] Generally, children from lower socioeconomic status are more likely to develop asthma with a greater symptom, though the definition of SES varied from study to study. It has been suggested that low income, low parental education level, higher likelihood of parental smoking in this group, and less likely to send children to day care centres were some of the causative factors.[14,15] However, it seems that the setting of these studies and the geographical location plays a role as results from Singapore, China, and South Africa showed an opposite effect.[16-18] Other factors such as decreasing family size, small sibling number, and birth order have been found to be associated with asthma likely in some sort of indirect way that is associated with microbial exposure in infancy and childhood.[19,20] In other words, children who are exposed to more microbes are less likely to develop asthma. This is further seen in those living on farms. The PARSIFAL and GABRIELA studies found correlation between farming environments and increased microbial burden associated with decreased asthma outcomes.[21,22] They found that farm living was associated with a greater number of gram-positive and gram-negative microbes, as well as bacterial and fungal toxins in mattresses.[21,23,24] A potential explanation for this is a connection between farming exposure and the activation of Toll-like receptor recognition signalling pathways which enhance the Th1 response, as opposed to the Th2.[24,25]

Exposure to air pollution has been shown to both exacerbate childhood asthma and also play a role in asthma development.[26,27] Studies looking at air pollution exposure during pregnancy and early life have shown that it is associated with changes in immune responses as well as structural remodelling of the lung parenchyma.[28,29] Early childhood exposure to traffic related air pollution, of which the main constituents are NO_2, $PM_{2.5}$, and PM_{10}[30], has been shown to increase the levels of proinflammatory cytokines associated with asthma, as well as elevated expression of the Clca3[30] gene that leads to mucous cell metaplasia and hyper-reactivity owing to the development of episodic recurrent airway obstruction.[31-33] Traffic related air pollution has also been associated with the transient and persistent wheeze phenotypes.[34]

Within the scope of this PhD, it is difficult to comment on the effects of these elements on the results presented as the data was not there to analyse.

7.7.3 Systematic Review

Using data-driven methods to ascertain patterns of asthma symptoms in childhood has led to large heterogeneity in results published by various individual studies. This disparity can be due to biases in the way the studies are conducted or designed, use of methodology that hasn't yet been standardised, misinterpretation or misrepresentation of results, or possibly the inherent heterogeneous nature of the disease that is yet to be fully understood. As such, it can be difficult to ascertain which results are the most reliable.

A systematic review aims to identify, evaluate, and summarise the findings of all studies relevant to the proposed topic, in order to report the best available research evidence that could aid in public policy, guideline creation/modification, and to guide future research based on identifying gaps in knowledge. Systematic reviews can be of interventions (randomised control trials) or observations (case studies, population cohorts, birth cohorts, etc). If done for interventions, it can provide high level evidence on the effectiveness of that intervention. The method of undertaking a systematic review adheres to a strict scientific design based on explicit, pre-specified and reproducible methods so that conclusions can be justified. They can often, but not always, include a meta-analysis which utilises statistical methodology to synthesize the results from different studies into a single quantitative estimate or summary effect size.

A systematic review differs from a narrative or rapid review in that rapid reviews are mainly descriptive, do not involve a protocol driven systematic search of the literature, usually have only one author choosing a subset of studies based on some less defined criteria. Although informative, they can often include an element of selection bias.[35] The process for undertaking a systematic review involves 7 steps which are summarised below.

Identify and formulate a review question

The first step is to formulate a review question that could be based on some background knowledge and use of an expert group opinion would be beneficial. A search of the DARE (Database of Abstracts of Reviews of Effects) and/or the CDSR (Cochrane Database of Systematic Reviews) databases can identify if such a review has previously been undertaken and to ensure a new review is justified.

Defining inclusion and exclusion criteria

The most common method of identifying inclusion/exclusion criteria is using the PICOS (population, intervention, comparison, outcomes, study design) method. Furthermore, a decision needs to be made on what type of studies will be included, how old the studies can be, any language restrictions, and whether using published/unpublished work. In the case of this PhD, no randomised control studies would be included as they are not available.

Population

- Who does this pertain to?
- Deciding on a minimum number of participants per group
- Deciding on the age, gender, ethnicity, deprivation status, co-morbidities, socioeconomic status, and geographical area (in the case of chapter 2 and 3 of this PhD, all children under age 18 were chosen. There were no limitations on the other demographic factors)
- The chosen population should have relevance to the population where the review findings will be applicable
- The severity of the disease in the population

Intervention

- What is the intervention or cause?
- Nature and setting of the intervention given

Comparison

- For comparative studies, then a decision needs to be made on what elements are considered as comparisons (ie. Methodology, data input, choice of population, etc)
- Is there something to compare the intervention to?

Outcomes

- A clearly defined set of relevant outcomes with corresponding justification
- In the case of chapter 2 and 3 of this PhD, primary outcomes are wheeze phenotypes by age, triggers and classification of asthma severity.

Study Design

- Randomised control trials, non-randomised control trials, observational studies (cohort, case-control, case series)
- Selected or unselected cohort
- Cross-sectional or longitudinal

Developing search strategy

Search strategies usually start with a list of key terms (MeSH) related to each of the PICOS components. It is necessary to develop a search strategy that can balance sensitivity (high proportion of relevant studies) and specificity (low proportion of irrelevant studies). Having at least one more person would be beneficial so that a consensus can be achieved. The MeSH list is a dynamic process where terms can be added whilst going through the literature. Common databases used for searching are: Medline, Ovid, PsychNet, Cochrane Library, Scopus.

Study selection

At least two reviewers (for increased inter-rater reliability) should screen titles/abstracts that arise from the search criteria and choose those that appear to meet the inclusion criteria. This leads to the creation of a list of papers to be read in more detail. If there is a lack of agreement on certain articles, then a third person can be introduced.

Data extraction

This should be done by at least two reviewers whereby a table is created to organise relevant information extracted from each study (i.e. Authors, publication year, population size, age range, study design, outcomes, included/excluded).

Assessing study quality

Assessing study quality is still rather subjective. Although there are now some standardised guides on the reporting of studies, there is no consensus on what constitutes good quality. For randomised control trials, the CONSORT (Consolidated Standards of Reporting Trials) Statement (http://www.consort-statement.org) is used. Guidelines used for reporting different study designs can be either EQUATOR (http://www.equator-network.org) or STROBE (http://www.strobe-statement.org). With regards to study quality, key aspects assessed are: 1) appropriateness of study design to research objective, 2) risk of bias, 3) choice of outcome measure (s), 4) statistical methodology, 5) quality of reporting, 6) quality of intervention, 7) generalisability.

Analysing and interpreting the results

The main calculable measure is the effect size for meta analyses. Effect sizes are represented as a value with a 95% confidence interval. A heterogeneity value can also be calculated to indicate whether individual studies are similar enough to compare. The final step would then be to summarise these results and provide recommendations for further clinical work or research.

7.7.4 Information Governance and Data Ethics

Information governance is a framework that allows researchers to handle personal and sensitive information in a transparent manner while upholding confidentiality and security. It mandates that information is held securely and confidentially, obtained in a fair manner, recorded accurately and reliable, used effectively and ethically, and shared appropriately and lawfully (https://www.hra.nhs.uk). With the introduction of the General Data Protection Regulation (GDPR), emphasis is placed on how to handle personal data. One of the key elements of this is consent, which is sought for participation in research studies, and participants need to be informed of how their information will be used. Consent can be withdrawn at any point. Patient identifiable data is of particular importance as it can reveal the identity of the participant. Patient identifiable data includes name, address, date of birth, hospital/NHS number. The data should then either be fully anonymised (information which cannot be reasonably used for identification) or pseudonymised (a unique code given to each participant that only those who have access to original data would be able to identify while the rest see it as anonymised). Subjects whose data was collected must be informed of the purpose of the data collection. If the purpose of the data changes, the subjects must be informed. However, an exception for research purposes exists whereby there are organisational measures in place that respect the principle of data minimisation (by removing patient identifiable data), or if giving information to data subjects would be impossible or would involve 'disproportionate effort'.

Health Research Authority and Research Ethics Committee approval

HRA approval includes the assessment of governance and legal compliance as well as the ethical opinion of the Research Ethics Committee. All projects classed as research require approval, and this includes: clinical trials of an investigational medicinal product, clinical investigation of a medical device, randomised control trial, basic study involving procedures with human participants, studies with questionnaires or interviews, studies with human tissue samples, and studies limited to working with data.

7.7.5 References

1. Williams LK, Pladevall M, Xi H, et al. Relationship between adherence to inhaled corticosteroids and poor outcomes among adults with asthma. *J Allergy Clin Immunol* 2004; **114**(6): 1288-93.
2. Murphy AC, Proeschal A, Brightling CE, et al. The relationship between clinical outcomes and medication adherence in difficult-to-control asthma. *Thorax* 2012; **67**(8): 751-3.
3. Gamble J, Stevenson M, McClean E, Heaney LG. The prevalence of nonadherence in difficult asthma. *Am J Respir Crit Care Med* 2009; **180**(9): 817-22.
4. Klok T, Kaptein AA, Brand PLP. Non-adherence in children with asthma reviewed: The need for improvement of asthma care and medical education. *Pediatr Allergy Immunol* 2015; **26**(3): 197-205.
5. Beck-Ripp J, Griese M, Arenz S, Koring C, Pasqualoni B, Bufler P. Changes of exhaled nitric oxide during steroid treatment of childhood asthma. *Eur Respir J* 2002; **19**(6): 1015-9.
6. Klok T, Brand PLP. Can exhaled nitric oxide fraction predict adherence to inhaled corticosteroids in atopic and nonatopic children with asthma? *J Allergy Clin Immunol Pract* 2017; **5**(2): 521-2.
7. Koster ES, Raaijmakers JA, Vijverberg SJ, Maitland-van der Zee AH. Inhaled corticosteroid adherence in paediatric patients: the PACMAN cohort study. *Pharmacoepidemiol Drug Saf* 2011; **20**(10): 1064-72.
8. Almqvist C, Pershagen G, Wickman M. Low socioeconomic status as a risk factor for asthma, rhinitis and sensitization at 4 years in a birth cohort. *Clin Exp Allergy* 2005; **35**(5): 612-8.
9. Gong T, Lundholm C, Rejno G, Mood C, Langstrom N, Almqvist C. Parental socioeconomic status, childhood asthma and medication use--a population-based study. *PLoS One* 2014; **9**(9): e106579.
10. Lindbaek M, Wefring KW, Grangard E, Ovsthus K. [Socioeconomic conditions and asthma in 4-5-year old children--a cohort study in Vesthold]. *Tidsskr Nor Laegeforen* 2003; **123**(9): 1187-90.
11. Kozyrskyj AL, Kendall GE, Jacoby P, Sly PD, Zubrick SR. Association between socioeconomic status and the development of asthma: analyses of income trajectories. *Am J Public Health* 2010; **100**(3): 540-6.
12. Braback L, Hjern A, Rasmussen F. Social class in asthma and allergic rhinitis: a national cohort study over three decades. *Eur Respir J* 2005; **26**(6): 1064-8.
13. Mielck A, Reitmeir P, Wjst M. Severity of childhood asthma by socioeconomic status. *Int J Epidemiol* 1996; **25**(2): 388-93.
14. Stein RT, Holberg CJ, Sherrill D, et al. Influence of parental smoking on respiratory symptoms during the first decade of life: the Tucson Children's Respiratory Study. *Am J Epidemiol* 1999; **149**(11): 1030-7.
15. Horwood LJ, Fergusson DM, Shannon FT. Social and familial factors in the development of early childhood asthma. *Pediatrics* 1985; **75**(5): 859-68.
16. Lai CK, Douglass C, Ho SS, et al. Asthma epidemiology in the Far East. *Clin Exp Allergy* 1996; **26**(1): 5-12.
17. Goh DY, Chew FT, Quek SC, Lee BW. Prevalence and severity of asthma, rhinitis, and eczema in Singapore schoolchildren. *Arch Dis Child* 1996; **74**(2): 131-5.

18. Poyser MA, Nelson H, Ehrlich RI, et al. Socioeconomic deprivation and asthma prevalence and severity in young adolescents. *Eur Respir J* 2002; **19**(5): 892-8.

19. Ball TM, Castro-Rodriguez JA, Griffith KA, Holberg CJ, Martinez FD, Wright AL. Siblings, day-care attendance, and the risk of asthma and wheezing during childhood. *N Engl J Med* 2000; **343**(8): 538-43.

20. Strachan DP. Family size, infection and atopy: the first decade of the "hygiene hypothesis". *Thorax* 2000; **55 Suppl 1**: S2-10.

21. Ege MJ, Mayer M, Normand AC, et al. Exposure to environmental microorganisms and childhood asthma. *N Engl J Med* 2011; **364**(8): 701-9.

22. Braun-Fahrlander C, Riedler J, Herz U, et al. Environmental exposure to endotoxin and its relation to asthma in school-age children. *N Engl J Med* 2002; **347**(12): 869-77.

23. Alfven T, Braun-Fahrlander C, Brunekreef B, et al. Allergic diseases and atopic sensitization in children related to farming and anthroposophic lifestyle--the PARSIFAL study. *Allergy* 2006; **61**(4): 414-21.

24. von Mutius E, Radon K. Living on a farm: impact on asthma induction and clinical course. *Immunol Allergy Clin North Am* 2008; **28**(3): 631-47, ix-x.

25. Lauener RP, Birchler T, Adamski J, et al. Expression of CD14 and Toll-like receptor 2 in farmers' and non-farmers' children. *Lancet* 2002; **360**(9331): 465-6.

26. Tzivian L. Outdoor air pollution and asthma in children. *J Asthma* 2011; **48**(5): 470-81.

27. Gowers AM, Cullinan P, Ayres JG, et al. Does outdoor air pollution induce new cases of asthma? Biological plausibility and evidence; a review. *Respirology* 2012; **17**(6): 887-98.

28. Peden DB. Development of atopy and asthma: candidate environmental influences and important periods of exposure. *Environ Health Perspect* 2000; **108 Suppl 3**: 475-82.

29. Kajekar R. Environmental factors and developmental outcomes in the lung. *Pharmacol Ther* 2007; **114**(2): 129-45.

30. Yang J, Chen Y, Yu Z, Ding H, Ma Z. Changes in gene expression in lungs of mice exposed to traffic-related air pollution. *Mol Cell Probes* 2018; **39**: 33-40.

31. Hajat A, Allison M, Diez-Roux AV, et al. Long-term exposure to air pollution and markers of inflammation, coagulation, and endothelial activation: a repeat-measures analysis in the Multi-Ethnic Study of Atherosclerosis (MESA). *Epidemiology* 2015; **26**(3): 310-20.

32. Rincon M, Irvin CG. Role of IL-6 in asthma and other inflammatory pulmonary diseases. *Int J Biol Sci* 2012; **8**(9): 1281-90.

33. Hoffmann B, Moebus S, Dragano N, et al. Chronic residential exposure to particulate matter air pollution and systemic inflammatory markers. *Environ Health Perspect* 2009; **117**(8): 1302-8.

34. Gehring U, Wijga AH, Brauer M, et al. Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life. *Am J Respir Crit Care Med* 2010; **181**(6): 596-603.

35. Petticrew M. Why certain systematic reviews reach uncertain conclusions. *BMJ* 2003; **326**(7392): 756-8.