Curr Allergy Asthma Rep (2015) 15: 38
DOI 10.1007/s11882-015-0542-0

IMMUNOLOGIC/DIAGNOSTIC TESTS IN ALLERGY (M CHAPMAN AND A POMÉS, SECTION EDITORS)

Distinguishing Asthma Phenotypes Using Machine Learning Approaches

Rebecca Howard[1] · **Magnus Rattray**[2] · **Mattia Prosperi**[1,3] · **Adnan Custovic**[4]

Published online: 5 July 2015

Abstract Asthma is not a single disease, but an umbrella term for a number of distinct diseases, each of which are caused by a distinct underlying pathophysiological mechanism. These discrete disease entities are often labelled as 'asthma endotypes'. The discovery of different asthma subtypes has moved from subjective approaches in which putative phenotypes are assigned by experts to data-driven ones which incorporate machine learning. This review focuses on the methodological developments of one such machine learning technique—latent class analysis—and how it has contributed to distinguishing asthma and wheezing subtypes in childhood. It also gives a clinical perspective, presenting the findings of studies from the past 5 years that used this approach. The identification of true asthma endotypes may be a crucial step towards understanding their distinct pathophysiological mechanisms, which could ultimately lead to more precise prevention strategies, identification of novel therapeutic targets and the development of effective personalized therapies.

Keywords Asthma · Allergy · Endotypes · Phenotypes · Machine learning · Childhood asthma · Latent class analysis

Supported in part by the UK Medical Research Council (MRC) Grant MR/K002449/1 and the MRC Health eResearch Centre (HeRC) grant MR/K006665/1

This article is part of the Topical Collection on *Immunologic/Diagnostic Tests in Allergy*

Rebecca Howard, Magnus Rattray, Mattia Prosperi and Adnan Custovic contributed equally to this work.

✉ Adnan Custovic
adnan.custovic@manchester.ac.uk

[1] Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, UK

[2] Faculty of Life Sciences, University of Manchester, Manchester, UK

[3] University of Florida, Gainesville, FL, USA

[4] Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester and University Hospital of South Manchester, Manchester M23 9LT, UK

Introduction

Asthma is increasingly recognized as a heterogeneous condition [1], an umbrella diagnosis for several diseases which present with common symptoms such as wheeze and cough, but differ in their aetiology, pathogenesis and responses to treatment (Fig. 1) [2–9]. These variants of asthma have been described as 'asthma endotypes' [10, 11]. Unlike phenotypes, which are defined by sharing similar observable characteristics, endotypes may be defined as subtypes of a condition with overlapping clinical symptoms, but each being caused by a distinct underlying pathophysiological mechanism [10] (Fig. 1).

At this moment, 'asthma endotype' is predominantly a hypothetical construct which has a potential value in helping us to uncover the mechanisms underlying different diseases in the 'asthma syndrome' [12]. Unravelling unique mechanisms for each asthma endotype may improve our understanding of the natural history of these diseases and ultimately could lead to more precise (possibly mechanism-specific) prevention strategies and may be crucial for the development of more effective personalized therapies and stratified health care [7, 8, 12, 13•]. For this to be feasible, amongst patients with the diagnostic label of 'asthma', it will be necessary to distinguish between different endotypes more precisely and in an unbiased way, as opposed to the currently prevailing classifications based on simple clinical phenotypic characteristics which usually focus on a single dimension of the disease (such as

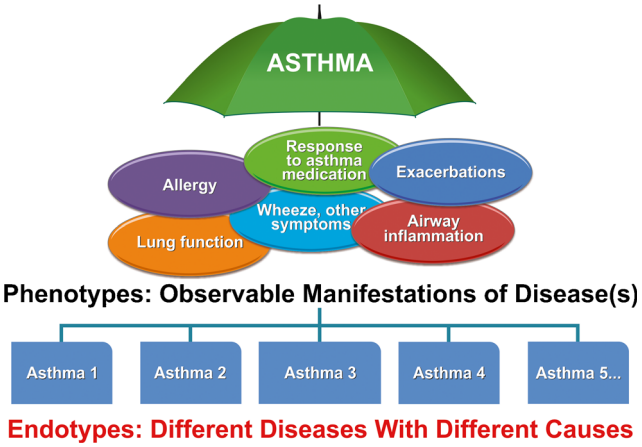

Figure 1 Asthma: an umbrella diagnosis which comprises multiple diseases with distinct mechanisms

eosinophilic or neutrophilic inflammation). One of the obstacles to this approach is that the information domains from which endotypes should be identified are not well defined. This, in conjunction with the persistence of different definitions of asthma in the medical literature affects the performance of prediction models [12, 14].

Investigator-Imposed vs. Data-Driven Approaches to Subtyping Asthma

The current approaches used to classify patients into asthma sub-groups can be split into two main types: subjective (i.e. investigator-imposed) and data-driven. The former is often referred to as being hypothesis-driven, and the latter as hypothesis-generating. In most subjective approaches, an investigator (usually an expert in the field) reviews the patterns of change in an individual's symptoms, triggers, pathology or airway obstruction, and then classifies patients into different 'phenotypes'. An example of this approach which withstood the test of time is a seminal publication from the Tucson Children's Respiratory Study, which described phenotypes of wheezing illness in pre-school children based on clinical assessment of whether a child had wheezed in the previous 12 months at ages 3 and 6 years [15]. The children with wheezing were assigned to three phenotypes: transient early wheezers, late-onset wheezers and persistent wheezers. However, by using this approach, one cannot estimate the uncertainty of phenotype classification. The research in this area has therefore moved on from using classifications predetermined by experts towards data-driven methodologies. Such approaches incorporate statistical learning techniques, facilitating the exploration of high dimensional clinical data sets. Here, we review the last 5 years of developments of latent variable modelling techniques—specifically, the latent class analysis (LCA) [16]—through which the dimensionality of the data sets can be reduced and variables can be grouped into patterns. We will describe the main cohorts used in cited studies and discuss issues related to the analytical approaches used in different studies (such as model definition and parameter estimation, model selection and class assignment) and the advantages and disadvantages of cross-sectional vs. longitudinal LCA approaches.

Background on Latent Class Analysis

Model Definition and Parameter Estimation: Which Classes?

Latent class analysis aims to fit a probabilistic model to data containing observable variables such as asthma symptoms (e.g. wheeze and/or cough) or atopic sensitization (e.g. serum specific IgE levels and skin prick tests): the observed variables are considered to be imperfect indicators of a set of unobserved latent variables. The assumption underpinning LCA is that all associations amongst the observed variables are due to the unobserved latent classes. The probabilistic model used for LCA is often referred to as a mixture model, because the probability of the observed data is a weighted sum, or mixture, of the data probability for each latent class. The weights of the mixture are the prior probabilities of each latent class, i.e. the chance to observe that class in the overall population. Latent classes are derived entirely from the observed data in an unsupervised manner [17, 18]. The subjects assigned to each class will be similar to each other according to the descriptor variables used, and the latent classes should correspond to clusters of similar subjects. LCA has been shown to be appropriate for modelling data on the occurrence or absence of symptoms in heterogeneous diseases such as asthma [3, 19–21], with the assumption that the co-occurrence of symptoms within the resulting latent classes is the consequence of unique disease-specific mechanisms, and that therefore the latent classes may be regarded as *bona fide* asthma endotypes.

In order to produce the latent classes, the model estimates two important quantities:

1. Conditional probability of each variable's response within each class.
2. Posterior probabilities of class membership for each subject given their response history.

Parameter estimation in the LCA (e.g. weights of the mixtures or average variable values in a class) can be done via different methods, one of the most popular being the expectation-maximization (EM) algorithm. In EM, two distinct steps (called E- and M-step) iteratively find the best parameter set by using a current (best guess) estimation on unseen data and improving the estimate by recalculating its likelihood (i.e. how well the model

fits/explains data points) on observed data, until the two estimates do not converge to the same value. It is typically assumed that each class has a characteristic distribution of the key variables modelled by some parameterized density function [22]. Often, the choice of class density function is restricted by the need for the EM optimization scheme to be tractable, e.g. class distributions from the exponential family are often chosen since the M-step of the EM algorithm is exactly tractable in this case. An alternative estimation procedure that has been used is the approximate Bayesian approach of variational message passing [23, 24, 25•].

The models typically used for LCA assume conditional independence of observed variables within each latent class, which is a strong assumption. Original data may be reduced into independent components (thus reducing the original number of variables) using techniques such as principal component analysis (PCA), exploratory factor analysis or multiple correspondence analysis [26], as variables representing the same dimensions are thought to be likely to be dependent within all identified latent classes. These variables are also typically required to be categorical, not continuous. Additionally, covariates such as sex effects can be (and have been) estimated [27, 28•].

Model Selection: How Many Classes?

The EM-algorithm can be used to learn the model parameters. However, this leaves the problem of choosing the optimal number of latent classes. This model selection problem is much more challenging than parameter estimation. It cannot be solved by simply maximizing the data likelihood: it is always true that a more complex model (i.e. more parameters) can achieve the same or higher likelihood than a simpler model (i.e. less parameters). Therefore, it is necessary to account for such model complexity (the number of free parameters) when selecting the optimal model.

There is not yet a single agreed method that determines the optimal number of latent classes for a model [29], although some methods are better suited for specific cases (or more popular) than others. Often, a variety of different model selection procedures are used, with their results compared and interpreted to help determine the model with the best number of classes. The most popular method used amongst the literature we reviewed was the Bayesian Information Criterion (BIC) [20, 21, 26, 27, 28•, 30••, 31, 32••, 33•, 34–37] along with some variations on its original formulation [28•, 38]. The function combines a model's log-likelihood value which is penalized by the number of parameters [39]—namely, the model log-likelihood is penalized by subtracting a quantity proportional to the number of parameters times the logarithm of the sample size—where the penalty can be interpreted in a Bayesian fashion. Complex models are penalized to avoid over-fitting, and parsimony is rewarded [40]. The model with the lowest BIC value is considered the best-fitting model and is usually selected [32••]. Its solution can depend on the sample size and the starting value. To account for the former, the algorithm can be adjusted [37]. To negate the latter, analyses are often rerun with many different starting values to confirm stable solutions and to avoid local maxima [37, 41]. The BIC is currently regarded as one of the most efficient in-sample estimators [31, 38] that does not require an external test set.

Another popular method is the Akaike Information Criterion (AIC) [26, 27, 28•, 42••], which is very similar to the BIC. The former uses a smaller penalty term for the number of parameters in a model [43]—i.e. only the number of parameters, without accounting for the sample size. An adjusted version was also used [28•, 38]. As with the BIC, a lower value indicates a superior model [16].

The next most popular methods are likelihood ratio tests [44] (bootstrap [21, 28•, 30••, 31, 33•, 34, 35], Lo-Mendell-Rubin [28•, 45•, 46] and chi-squared [36]), all of which are parametric [22, 38]. These test an improvement in fit between models with n vs. $n+1$ classes, resulting in a fit index [41]. For the tests that assume a chi-squared distribution, their results may be affected by sample sizes as the test statistic follows the distribution asymptotically. The bootstrap likelihood ratio test instead constructs a distribution using parametric bootstrap approach and as such should be less affected by sample sizes [22, 31].

Other methods used to select the models used include entropy [21, 27, 28•, 42••, 45•], the Bezdek partition coefficient [26], confusion matrices [23] and a dissimilarity index [36]. More recently, Bayesian latent variable methods closely related to LCA have also been introduced which allow models to be selected using the Bayesian evidence (also referred to as the Marginal likelihood) [47].

The selected model's resulting classes are interpreted as distinct subtypes, characterized by the model variables that apply best to each class. Because of this, classes that share very similar model variable characteristics are interpreted as representing very similar phenotypes and so are often merged in favour of fewer classes [31].

Class Assignment: How Are (New) Subjects Assigned to the Latent Class?

For each subject, the probability of belonging to each of the latent classes is calculated. In a process called modal allocation, each subject is assigned to the latent class with the largest a posteriori probability of membership [19]. A classification is supported by high membership probabilities, indicating good separation between clusters. These classifications are validated by testing the reproducibility of these classes [26, 32••], sensitivity analyses [26] and the analysis of their association

with objective measurements considered to be relevant to asthma such as lung function (e.g. forced expiratory volume in 1 s [FEV_1]), and bronchial hyperresponsiveness (BHR) [31] using regression analyses [28•, 32••, 36, 45•, 48] and true- and false-positive rates from receiver operating characteristic curves [42••, 49].

Strengths and Limitations of Different LCA Approaches

A major advantage of LCA is that subjects are not absolutely assigned to a single class, but instead have probabilities of membership to various classes. Also, as the model selection criteria previously described assist in determining the optimal number of classes, LCA can be regarded as fairly objective in the sub-group sets identification [19, 50]. However, a degree of subjectivity is introduced into LCA as some a priori decisions need to be made, such as which variables will be included in the model, with the assumptions on the data values distributions, and on the class shapes [19].

LCA is particularly suitable for categorical input variables and can accommodate missing values when they are assumed to be missing at random [51], thus allowing the analysis of a whole sample. However, caution is required when using records containing missing values, as their use can pose a high risk of bias where missing values are correlated with clinical attributes. For example, in some studies, allergy tests are rarely performed in non-asthmatic children, and a lack of information about family history, environmental exposures and clinical features can also contribute to bias. There is also the risk of unrecorded episodes (e.g. of wheeze), which can lead to an underestimation of incidence rates, possibly resulting in imprecise class definitions [34]. Manifestation of biases such as these could result in misclassification of subjects.

Another LCA approach, longitudinal latent class analysis (LLCA) can temper these biases as it accounts for correlation between reports at different time points [30••, 51, 52]. This is possible because LLCA clusters individuals into classes alongside others that share similar longitudinal response patterns across discrete time points. Subjects with sporadic or incomplete reports are assigned to classes with less certainty that those with consistent reports across time or those who report patterns consistent with other subjects. However, LLCA sometimes requires responses to be collected at the same discrete time point in each subject. This has major implications for data collection. Ideally, every subject should be exactly the same age when each measurement is performed. However, in most epidemiological longitudinal studies, most measurements are not collected in this fashion, so the rounding of age is required, introducing measurement errors [33•]. Also, LLCA also does not allow the modelling of the effect of time-varying causative factors (such as environmental exposures, e.g. seasons) on the prevalence of a response such as wheeze.

These limitations can be overcome by using another, more flexible form of LCA, latent class growth analysis (LCGA) [41, 53]. Here, the variables need not be categorical, but can be continuous, removing the need to round ages or other numerical measurements. This allows trajectories of development (e.g. of wheeze or atopic sensitization) to be estimated as a continuous function of age. The effect of time-varying causative factors can now be included as well (for example the effect of common cold/flu season on the prevalence of wheezing) [30••]. This relatively novel method enables the investigation of associations between time-varying and time-invariant factors on response patterns and makes it easier to compare classes across different populations and to account for a variable number of repeated assessments. All of this can be particularly advantageous when studying the effects of repeated environmental exposures and their outcomes that fluctuate over time through childhood.

Software Implementations

LCA methods are implemented and are available in many statistical and machine learning tools such as—from the literature here reviewed—Mplus (Los Angeles, CA, USA) [28•, 35, 42••, 48, 54, 55], Latent GOLD (Statistical Innovations, Boston, USA) [31, 34, 56], R (http://www.r-project.org/), including its package poLCA (Emory University, Atlanta, GA, USA) [26, 57], Infer.NET (Microsoft, Cambridge, USA) [25•, 58], STATA (StataCorp, College Station, TX, USA) [32••, 40, 59, 60], SAS Statistical Software (SAS Institute, Cary, NC, USA) [61], including the PROC LCA [20, 27, 30••, 51, 52] and TRAJ procedures [30••, 61, 62], and Multimix (University of Waikato, Hamilton, New Zealand) [33•, 63].

Cohort Details and Phenotype Associations

Cohorts that have utilized data-driven approaches to analysis in the field of asthma and allergy within the last 5 years are listed in Table 1. They are a mixture of longitudinal and cross-sectional studies, with the 19 out of the 25 being birth cohorts (see Table 1). Sample sizes used in different analyses range from 201 to 11,632 participants. Variants of LCA performed upon these study populations include LLCA, LCGA and latent growth mixture modelling (LGMM) [54]. The number of resulting classes for each of these studies ranged from 3 to 8, with the patients allocated to each class based on characteristics such as wheezing [21, 26, 30••, 31, 32••, 33•, 34, 36, 42••, 45•, 64, 65, 66••], atopic status [20, 23, 25•, 26, 28•, 35, 36, 66••] and coughing [31, 33•]. Studies that based their subtypes upon other characteristics included that of Figueiredo et al.

Table 1 Features of main asthma cohorts. Cohort features including name, nationality, cohort type, the age of the patients enrolled, the sample sizes and latent class analysis (LCA) approaches used, the resulting number of latent classes identified and what each one's phenotype was based upon is listed. The variants LCA performed include longitudinal latent class analysis (LLCA), latent class growth analysis (LCGA) and latent growth mixture modelling (LGMM). The phenotypes described by Bochenek et al. were based upon sub-phenotypes of aspirin-exacerbated respiratory disease (AERD). The full names of the studies are as follows: Avon Longitudinal Study of Parents and Children (ALSPAC); Asthma Multicentre Infant Cohort Study (AMICS); Asthma Multicentre Infant Cohort Study-Menorca (AMICS-Menorca); Cohort from Barreto et al., 2007; Cohort from Bochenek et al., 2013; Childhood Asthma Prevention Study cohort (CAPS); Columbia Center for Children's Environmental Health (CCCEH); Danish Allergy Research Council study (DARC); European Community Respiratory Health Survey (ECRHSII); Epidemiological Study on the Genetics and Environment of Asthma (EGEA2); German Infant Nutritional Intervention Study (GINIplus); Isle of Wight cohort (IoW); International Study on Allergies and Asthma in Childhood—Spain—Phase II (ISAAC phase II); Leicester Respiratory Cohort; The Influences of Lifestyle-related factors on the Immune System and the Development of Allergies in Childhood study (LISA); Manchester Asthma and Allergy Study (MAAS); Melbourne Atopy Cohort Study (MACS); Multicenter Allergy Study (MAS); Millennium Cohort Study (MCS); Pollution and Asthma Risk: An Infant Study birth cohort (PARIS); Protection against Allergy—Study in Rural Environment (PASTURE); Prevention and Incidence of Asthma and Mite Allergy study (PIAMA); Prospective Study on the Influence of Perinatal factors on the Occurrence of Asthma and Allergies (PIPO); Sibilancias de Lactante y Asma de Mayor study (SLAM); Wayne County Health, Environment, Allergy and Asthma Longitudinal Study (WHEALS)

Cohort name	Cohort references	Nationality	Cohort type	Age		Analysis references	Sample size used	LCA approach	Number of classes	Phenotypes based upon:
ALSPAC	[3, 77]	British	Birth cohort	Children	0–11 years	[21]	5760	LLCA	6	Wheezing
						[65]	7247	LLCA	6	Wheezing
						[66••]	8665	LLCA	8	Wheezing, atopic status, eczema and rhinitis
AMICS	[78]	European	Birth cohort	Children	–	[54]	487	LGMM	3	Growth patterns
AMICS-Menorca	[54]	Spanish	Birth cohort	Children	–	[54]	485	LGMM	3	Growth patterns
Barreto et al., 2007	[79]	Brazilian	Population-based	Children	0–42 months	[48]	1127	LCA	3	Cytokine production
Bochenek et al., 2014	[27]	Polish	Population-based	Unspecified	–	[27]	201	LCA	4	AERD sub-phenotypes
CAPS	[80–82]	Australian	Birth cohort	Children	0–8 years	[28•]	578	LLCA	4	Atopic status
CCCEH	[83, 84]	Columbian	Birth cohort	Children	0–9 years	[30••]	689	LLCA and LCGA	4	Wheezing
DARC	[85]	Danish	Birth cohort	Children	0–15 years	[54]	572	LGMM	3	Growth patterns
ECRHSII	[86, 87]	European	Population-based	Adults	–	[20]	1895	LCA	4	Atopic status
EGEA2	[88, 89]	French	Case–control and family-based study	Adults	–	[20]	641	LCA	4	Atopic status
GINIplus	[59, 90]	German	Birth cohort	Children	0–6 years	[54]	3739	LGMM	3	Growth patterns
IoW	[73–76]	British	Birth cohort	Children	0–18 years	[25•]	Unspecified	LCA	5	Atopic status
ISAAC phase II	[91, 92]	Spanish	Population-based	Children	8–12 years	[31]	>4000	LCA	7	Wheezing and coughing
Leicester respiratory Cohort	[93]	British	Population-based	Children	0–5 years	[33•]	319	LCA	5	Wheezing and coughing
LISA	[94]	German	Birth cohort	Children	0–6 years	[54]	3097	LGMM	3	Growth patterns
MAAS	[95–97]	British	Birth cohort	Children	0–11 years	[32••]	1184	LLCA	5	Wheezing
						[23]	1053	LCA	5	Atopic status

Table 1 (continued)

Cohort name	Cohort references	Nationality	Cohort type	Age		Analysis references	Sample size used	LCA approach	Number of classes	Phenotypes based upon:
						[25•]	1053	LCA	5	Atopic status
						[66••]	1136	LLCA	8	Wheezing, atopic status, eczema and rhinitis
MACS	[98]	Australian	Birth cohort	Children	0–7 years	[64]	620	LCA	5	Wheezing
MAS	[99]	German	Birth cohort	Children	0-(Ongoing)	[54]	1314	LGMM	3	Growth patterns
MCS	[100]	British	Birth cohort	Children	0–8 years	[36]	11,632	LLCA	4	Wheezing and atopic status
PARIS	[101]	French	Birth cohort	Children	0–18 months	[26]	1831	LCA	4	Wheezing and atopic status
PASTURE	[102]	European	Birth cohort	Children	0–6 years	[42••]	1133	LCA	5	Wheezing
PIAMA	[103]	Dutch	Birth cohort	Children	0–8 years	[21]	2810	LLCA	5	Wheezing
						[65]	2007	LLCA	5	Wheezing
						[45•]	2728	LLCA	5	Wheezing
						[54]	1128	LGMM	3	Growth patterns
PIPO	[104]	Belgian	Birth cohort	Children	0–1 years	[54]	810	LGMM	3	Growth patterns
SLAM	[34]	Spanish	Birth cohort	Children	0–6 years	[34]	3739	LCA	4	Wheezing
WHEALS	[105]	U.S.A.	Birth cohort	Children	0–2 years	[35]	1187	LCA	4	Atopic status

[48], who based theirs upon cytokine production, denoting a burden of infection; Bochenek et al. [27] who based theirs upon sub-phenotypes of a recognized sub-type of asthma—aspirin-exacerbated respiratory disease; Rzehak et al. [54] who based theirs upon body mass index trajectories; and Belgrave et al. [66••] who included eczema, wheeze and rhinitis.

The resulting sub-groups (classes) were then often associated with clinical features such as atopy, physician-diagnosed asthma and fractional exhaled nitric oxide (FeNO) [20, 23, 25•, 26, 28•, 32••, 36, 67, 68]. For example, FeNO—a surrogate biomarker of the degree of eosinophilic airway inflammation [69, 70]—measured at age 8 years in the prevention and incidence of asthma and mite allergy (PIAMA) cohort, was found to be different amongst wheezing phenotypes, but only in atopic children, fuelling speculation that the pathophysiology of wheezing phenotypes differs between atopic and non-atopic children, and that they are the result of differing endotypes [67].

Wheezing is the feature that the analyses reviewed here most commonly used to derive classes (which are often referred to as 'wheeze phenotypes'), some (but not all) of which are similar across different studies and analyses. Types of wheeze phenotypes that have been identified across the studies include early life wheeze (transient and prolonged), late-onset wheeze and persistent wheeze (controlled and troublesome) [7]. For the early life wheeze phenotypes, a prolonged early wheeze (PEW) was identified in the Avon Longitudinal Study of Parents and Children (ALSPAC) study [21], but it was not found in the PIAMA cohort in the same analysis. However, the latter cohort's transient early wheeze (TEW)—also identified in the ALSPAC cohort—appeared to be a combination of ALSPAC's PEW and TEW classes, both in terms of size and of the prevalence of wheeze over time. Children in these classes were found to have diminished lung function at age 6–8 years (i.e. after the wheeze had resolved) compared to those who had never wheezed in both cohorts [21]. These phenotypes of early childhood wheezing were not associated with allergic sensitization, eczema or rhinitis, and it has since been confirmed that the developmental profiles of eczema, wheeze and rhinitis are indeed heterogeneous [66••].

The late-onset wheeze phenotype is generally characterized as wheeze which starts after the age of 3 years which then persists into later childhood. Studies have mixed reports with respect to the association of late-onset wheeze with secondary asthma phenotypes such as lung function and bronchial hyperresponsiveness [7]. For example, ALSPAC, PIAMA and MAAS all found that children in this subgroup are significantly more likely to have bronchial hyperresponsiveness [3, 21, 32••], but only MAAS and ALSPAC found significant associations of late-onset wheeze with lung function impairment at the age of 6 years [3, 71].

Lastly, the persistent wheeze phenotypes have been characterized by diminished lung function by school age in all cohorts which assessed their association [7]. In contrast to other studies which described a single persistent wheeze class, in the MAAS birth cohort, the children with persistent wheeze fell into two distinct classes: persistent controlled wheeze (PCW) and persistent troublesome wheeze (PTW) [32••]. The PTW group had worse lung function and more reactive airways than all other groups, including PCW [32••]. However, it is worth noting that this analysis utilized information on wheezing derived from two different sources—parentally reported and physician-confirmed—likely enabling a more precise allocation into sub-groups.

Most studies reported a strong association between persistent wheezing and atopy, with more than 50 % of persistent wheezers having been found to be atopic [72]. However, atopy is not a feature that is unique to this class, and it is also present in other wheeze classes, and amongst children who have never wheezed. Thus, on its own atopic status is not a good discriminator of wheeze class. Several studies hypothesized that similar to wheezing illness, atopy may also comprise of several distinct subtypes. For example, Herr et al. identified three distinct atopic phenotypes regarding atopy within the first 18 months of life amongst participants in the PARIS cohort [26]. In the analysis spanning the first 8 years of life, Simpson et al. have identified different structure within the MAAS data, with the optimal model containing five classes (early sensitization to multiple allergen sources, late sensitization to multiple allergen sources, mite, non-dust mite and no latent vulnerability) [23]. The atopic class of early sensitization to multiple allergen sources was associated with persistent wheeze phenotypes and was very strongly associated with physician-diagnosed asthma in the school age (odds ratio ~30) [23]. Lazic et al. extended this work by including newly available skin test and IgE data from age 11 years from the MAAS cohort and that of the Isle of Wight cohort [25•, 73–76]. Very similar five-class models emerged across the two cohorts, suggesting that these atopy classes were stable across time and different populations. In both cohorts, children in the class with sensitivity to a wide variety of allergens were considerably more likely to have asthma compared to all other classes [25•]. The children in this class (comprising approximately one quarter of children defined as atopic using the standard definition) across both cohorts had significantly poorer lung function, most reactive airways, highest eNO and most hospital admissions for asthma. Of note, the associations between asthma presence and severity and conventionally defined atopy were much weaker. These results indicate that there is a latent heterogeneity in atopy, similar to that found in asthma/wheezing illness [23, 25•]. Because of this, attempting to define atopy as a dichotomous trait could well be an oversimplification, much as it would be to define childhood wheezing in such a fashion.

Further research will be necessary in order to replicate different asthma, wheeze and atopy subtypes across independent

cohorts, to assess their stability over time and to confirm the existence of distinct pathophysiological mechanisms underpinning each sub-type. Since unique pathophysiological mechanisms for the subtypes of wheezing illness and atopy identified so far using machine learning approaches have not as yet been elucidated, these cannot be considered as 'true endotypes' but are mostly hypothetical constructs to facilitate further research in this area. The identification of underlying biology and real endotypes may have major implications for effective and precise asthma prevention, treatment and management strategies, as it is anticipated that the different groups may respond differently to the treatments currently offered.

Conclusions

A distinct set of heterogeneous diseases with the diagnostic label of asthma may potentially be identified using data-driven, computational techniques such as latent class analysis. Such techniques disambiguate the complex patterns of symptoms shared by these different diseases. This may be a first step towards elucidation and better understanding of their distinct underlying pathophysiological mechanisms, which could facilitate the development of personalized mechanism-specific prevention strategies and more effective stratified therapies.

Compliance with Ethics Guidelines

Conflicts of Interest Adnan Custovic reports grants from Medical Research Council, grants from The JP Moulton Charitable Foundation, grants from North West Lung Research Centre Charity, grants from European Union 7th Framework Programme, grants from National Institute of Health Research, personal fees from Novartis, personal fees from Thermo Fisher, personal fees from AstraZeneca, personal fees from ALK and personal fees from GlaxoSmithKline.

Rebecca Howard, Magnus Rattray and Mattia Prosperi declare that they have no conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. A plea to abandon asthma as a disease concept. Lancet, 2006. 368(9537): p. 705.
2. Haldar P et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med. 2008;178(3):218–24.
3. 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):974–80.
4. Moore WC 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. Smith JA et al. 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 et al. International consensus on (ICON) pediatric asthma. Allergy. 2012;67(8):976–97.
7. Belgrave DC, Custovic A, Simpson A. Characterizing wheeze phenotypes to identify endotypes of childhood asthma, and the implications for future management. Expert Rev Clin Immunol. 2013;9(10):921–36.
8. 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.
9. Wenzel SE. Asthma: defining of the persistent adult phenotypes. Lancet. 2006;368(9537):804–13.
10. 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):355–60.
11. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. Lancet. 2008;372(9643):1107–19.
12. Custovic A. The Study Team for Early Life Asthma Research (STELAR) consortium "Asthma e-lab": team science bringing data, methods and investigators together. Thorax. 2015.
13. • Custovic A, Lazic N, Simpson A. Pediatric asthma and development of atopy. Curr Opin Allergy Clin Immunol. 2013;13(2):173–80. **Excellent review discussing the controversies surrounding the relationship between asthma and atopy, and between the endotypes they exhibit**.
14. Van Wonderen KE et al. Different definitions in childhood asthma: how dependable is the dependent variable? Eur Respir J. 2010;36(1):48–56.
15. Martinez FD et al. Asthma and wheezing in the first six years of life. The group health medical associates. N Engl J Med. 1995;332(3):133–8.
16. Hagenaars J, McCutcheon A. Applied latent class analysis. Cambridge: Cambridge University Press; 2002. p. 454.
17. Rabe-Hesketh S, Skrondal A. Classical latent variable models for medical research. Stat Methods Med Res. 2008;17(1):5–32.
18. Spycher BD, Minder CE, Kuehni CE. Multivariate modelling of responses to conditional items: new possibilities for latent class analysis. Stat Med. 2009;28(14):1927–39.
19. Spycher BD et al. Distinguishing phenotypes of childhood wheeze and cough using latent class analysis. Eur Respir J. 2008;31(5):974–81.
20. Siroux V et al. Identifying adult asthma phenotypes using a clustering approach. Eur Respir J. 2011;38(2):310–7.
21. Savenije OE et al. Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA. J Allergy Clin Immunol. 2011;127(6):1505–12.e14.
22. McLachlan G, Peel D. Finite mixture models. Wiley. 2000.
23. 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):1200–6.
24. Winn J, Bishop C. Variational message passing. J Mach Learn Res. 2005;6:661–94.
25. • Lazic N et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. Allergy. 2013;68(6):764–70. **Validation of the findings of a previous study [23], by showing that very similar atopy classes can be identified in an independent population**.

26. Herr M et al. Risk factors and characteristics of respiratory and allergic phenotypes in early childhood. J Allergy Clin Immunol. 2012;130(2):389–96.
27. Bochenek G et al. Certain subphenotypes of aspirin-exacerbated respiratory disease distinguished by latent class analysis. J Allergy Clin Immunol. 2014;133(1):98–103.
28.• Garden FL, Simpson JM, Marks GB. Atopy phenotypes in the Childhood Asthma Prevention Study (CAPS) cohort and the relationship with allergic disease: clinical mechanisms in allergic disease. Clinical and experimental allergy. J Br Soc Allergy Clin Immunol. 2013;43(6):633–41. **This study defined atopy phenotypes, and found a strong association of asthma with the mixed food and inhalant phenotype in early life, implying that food sensitization in that period may be a more significant indicator of subsequent asthma than previously thought**.
29. Storr CL et al. Empirically derived latent classes of tobacco dependence syndromes observed in recent-onset tobacco smokers: epidemiological evidence from a national probability sample survey. Nicotine Tob Res. 2004;6(3):533–45.
30.•• Chen Q et al. Using latent class growth analysis to identify childhood wheeze phenotypes in an urban birth cohort. Ann Allergy Asthma Immunol. 2012;108(5):311–5. **The first study to demonstrate the use of latent class growth analysis, which enables the modelling of time-invariant and time-varying (e.g. season) risk factors**.
31. Weinmayr G et al. Asthma phenotypes identified by latent class analysis in the ISAAC phase II Spain study. Clin Exp Allergy. 2013;43(2):223–32.
32.•• Belgrave DCM et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. J Allergy Clin Immunol. 2013;132(3): 575–583.e12. **This study jointly modelled observations of wheeze from both medical records and parental reports. The incorporation of the medical records meant the severity of each of the wheeze phenotypes identified was more accurately characterized, uncovering a novel class of persistent troublesome wheezing**.
33.• Spycher BD et al. Comparison of phenotypes of childhood wheeze and cough in 2 independent cohorts. J Allergy Clin Immunol. 2013;132(5):1058–67. **This study validates the findings of a previous study [19], by consistently identifying two wheeze phenotypes across two independent cohorts**.
34. Cano-Garcinuño A, Mora-Gandarillas I, S.S. Group. Wheezing phenotypes in young children: an historical cohort study. 2014;23(1):60–66.
35. Havstad S, et al. Atopic phenotypes identified with latent class analyses at age 2 years. J Allergy Clin Immunol. 2014.
36. Panico L et al. Asthma trajectories in early childhood: identifying modifiable factors. PLoS One. 2014;9(11):e111922.
37. Schwarz G. Estimating the dimension of a model. 1978:461–464.
38. Nylund KL, Asparoutiov T, Muthen BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct Equ Model Multidiscip J. 2007;14(4):535–69.
39. Burnham K, Anderson D. Multimodel inference: understanding AIC and BIC in model selection. Soc Methods Res. 2004;33:261–304.
40. Pickles A, Croudace T. Latent mixture models for multivariate and longitudinal outcomes. Stat Methods Med Res. 2010;19(3):271–89.
41. Jung T, Wickrama KAS. An introduction to latent class growth analysis and growth mixture modeling. Soc Personal Psychol Compass. 2008;2(1):302–17.
42.•• Depner M et al. Clinical and epidemiologic phenotypes of childhood asthma. Am J Respir Crit Care Med. 2014;189(2):129–38. **This study compared clinical definitions and LCA-derived definitions of asthma, and found that the phenotypes were well supported by the LCA analysis performed**.
43. Burnham K, Anderson D. Model selection and multimodel inference: a practical information-theoretic approach. 2 ed. Springer-Verlag. 2002.
44. Vuong Q. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989;57:307–33.
45.• Caudri D et al. Perinatal risk factors for wheezing phenotypes in the first 8 years of life. Clinical and experimental allergy. J Br Soc Allergy Clin Immunol. 2013;43(12):1395–405. **Assessed associations of perinatal factors with wheezing phenotypes**.
46. Lo Y, Mendell N, Rubin D. Testing the number of components in a normal mixture. Biometrika. 2001;88:767–78.
47. Barber D. Bayesian reasoning and machine learning. Cambridge: Cambridge University Press; 2012.
48. Figueiredo CA et al. Environmental conditions, immunologic phenotypes, atopy, and asthma: new evidence of how the hygiene hypothesis operates in Latin America. J Allergy Clin Immunol. 2013;131(4):1064–8.
49. Robin X et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
50. Magidson J, Vermunt J. Latent class models for clustering: a comparison with K-means. Can J Public Health. 2002;20:36–43.
51. Lanza ST et al. PROC CA: a SAS procedure for latent class analysis. Struct Equ Model Multidiscip J. 2007;14(4):671–94.
52. Lanza ST et al. PROC LCA & PROC LTA user's guide (version 1.2.7). Penn State: University Park, The Methodology Center; 2011.
53. Nagin D. Analyzing developmental trajectories: a semi-parametric, group-based approach. Psychol Methods. 1999;4:139–77.
54. Rzehak P et al. Body mass index trajectory classes and incident asthma in childhood: results from 8 European Birth Cohorts—a Global Allergy and Asthma European Network initiative. J Allergy Clin Immunol. 2013;131(6):1528–36.
55. Muthén LK, Muthén BO. Mplus user's guide. Muthén & Muthén: Los Angeles, CA.
56. Vermunt JK, Magidson J. LatentGOLD user's guide. Belmont: Statistical Innovations Inc; 2003.
57. Linzer DA, Lewis JB. PoLCA: an R package for polytomous variable latent class analysis. J Stat Softw. 2011;42(10):1–29.
58. Minka T et al. Infer.NET 2.6. Cambridge: Microsoft Research; 2014.
59. Rzehak P et al. Period-specific growth, overweight and modification by breastfeeding in the GINI and LISA birth cohorts up to age 6 years. Eur J Epidemiol. 2009;24(8):449–67.
60. Rabe-Hesketh S, Skrondal A, Pickles A. GLLAMM manual. U.C. Berkeley Division of Biostatistics. 2004.
61. Jones BL, Nagin DS, Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. Sociol Methods Res. 2001;29(3):374–93.
62. Jones BL, Nagin DS. Advances in group-based trajectory modeling and an SAS procedure for estimating them. Sociol Methods Res. 2007;35(4):542–71.
63. Hunt L, Jorgensen M. Mixture model clustering using the MULTIMIX program. Aust N Ze J Stat. 1999;41(2):154–71.
64. Lodge CJ et al. Childhood wheeze phenotypes show less than expected growth in FEV1 across adolescence. Am J Respir Crit Care Med. 2014;189(11):1351–8.
65. Savenije OE, 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.
66.•• Belgrave DC et al. Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies. PLoS Med. 2014;11(10):e1001748. **This study determined the individual development trajectories of wheeze, eczema and rhinitis in childhood and found that only a small proportion (~7%)**

followed a trajectory resembling "atopic march". Otherwise, the profiles were heterogeneous.
67. van der Valk RJP et al. Childhood wheezing phenotypes and FeNO in atopic children at age 8. Clin Exp Allergy J Br Soci Allergy Clin Immunol. 2012;42(9):1329–36.
68. Belgrave DCM et al. Trajectories of lung function during childhood. Am J Respir Crit Care Med. 2014;189(9):1101–9.
69. American Thoracic S, S. European Respiratory. ATS/ERS recommendations for standardized procedures for the online and offline measurement of exhaled lower respiratory nitric oxide and nasal nitric oxide. Am J Respir Crit Care Med. 2005;171(8):912–30.
70. Pijnenburg MW, De Jongste JC. Exhaled nitric oxide in childhood asthma: a review. Clin Exp Allergy. 2008;38(2):246–59.
71. Lowe LA et al. Wheeze phenotypes and lung function in preschool children. Am J Respir Crit Care Med. 2005;171(3):231–7.
72. Stein RT, Martinez FD. Asthma phenotypes in childhood: lessons from an epidemiological approach. Paediatr Respir Rev. 2004;5(2):155–61.
73. Sly PD et al. Early identification of atopy in the prediction of persistent asthma in children. Lancet. 2008;372(9643):1100–6.
74. Scott M, Kurukulaaratchy RJ, Arshad SH. Definitions are important and not all wheeze is asthma. Thorax. 2011;66(7):633. **author reply 633–4**.
75. Stanojevic S et al. Reference ranges for spirometry across all ages: a new approach. Am J Respir Crit Care Med. 2008;177(3):253–60.
76. Crapo RO 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.
77. Fraser A et al. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013;42(1):97–110.
78. Sunyer J et al. Maternal atopy and parity. Clin Exp Allergy. 2001;31(9):1352–5.
79. Barreto ML et al. Effect of city-wide sanitation programme on reduction in rate of childhood diarrhoea in northeast Brazil: assessment by two cohort studies. Lancet. 2007;370(9599):1622–8.
80. Marks GB et al. Prevention of asthma during the first 5 years of life: a randomized controlled trial. J Allergy Clin Immunol. 2006;118(1):53–61.
81. Toelle BG et al. Eight-year outcomes of the childhood asthma prevention study. J Allergy Clin Immunol. 2010;126(2):388–9–389.e1-3.
82. Mihrshahi S et al. The childhood asthma prevention study (CAPS): design and research protocol of a randomized trial for the primary prevention of asthma. Control Clin Trials. 2001;22(3):333–54.
83. Perera FP et al. Effects of transplacental exposure to environmental pollutants on birth outcomes in a multiethnic population. Environ Health Perspect. 2003;111(2):201–5.
84. Miller RL et al. Prenatal exposure, maternal sensitization, and sensitization in utero to indoor allergens in an inner-city cohort. Am J Respir Crit Care Med. 2001;164(6):995–1001.
85. Høst A et al. Clinical course of cow's milk protein allergy/intolerance and atopic diseases in childhood. Pediatr Allergy Immunol. 2002;13(s15):23–8.
86. Burney P. The changing prevalence of asthma? Thorax. 2002;57 Suppl 2:II36–9.
87. The European Community Respiratory Health Survey, I.I.S.C. The European Community Respiratory Health Survey II. Eur Respir J. 2002;20(5):1071–9.
88. Kauffmann F, Dizier MH. EGEA (Epidemiological study on the Genetics and Environment of Asthma, bronchial hyperresponsiveness and atopy)-design issues. Clin Exp Allergy. 1995;25(s2):19–22.
89. Kauffmann F et al. Epidemiological study of the genetics and environment of asthma, bronchial hyperresponsiveness, and atopy: phenotype issues. Am J Respir Crit Care Med. 1997;156(4 Pt 2):S123–9.
90. Berg AV et al. Impact of early feeding on childhood eczema: development after nutritional intervention compared with the natural course—the GINIplus study up to the age of 6 years. Clin Exp Allergy J Br Soc Allergy Clin Immunol. 2010;40(4):627–36.
91. Weiland SK et al. Phase II of the International Study of Asthma and Allergies in Childhood (ISAAC II): rationale and methods. Eur Respir J. 2004;24(3):406–12.
92. García-Marcos Álvarez L et al. International Study of Asthma and Allergies in Childhood (ISAAC) fase II: metodología y resultados de participación en España. Anales Pediatr. 2001;55(5):400–5.
93. Kuehni CE et al. Cohort profile: the Leicester respiratory cohorts. Int J Epidemiol. 2007;36(5):977–85.
94. Chen C-M et al. Longitudinal study on cat allergen exposure and the development of allergy in young children. J Allergy Clin Immunol. 2007;119(5):1148–55.
95. Lowe L et al. Specific airway resistance in 3-year-old children: a prospective cohort study. Lancet. 2002;359(9321):1904–8.
96. Nicolaou NC et al. Exhaled breath condensate pH and childhood asthma: unselected birth cohort study. Am J Respir Crit Care Med. 2006;174(3):254–9.
97. Custovic A et al. The national asthma campaign Manchester asthma and allergy study. Pediatr Allergy Immunol. 2002;13(s15):32–7.
98. Lowe AJ et al. Effect of a partially hydrolyzed whey infant formula at weaning on risk of allergic disease in high-risk children: a randomized controlled trial. J Allergy Clin Immunol. 2011;128(2):360–365.e4.
99. Bergmann RL et al. Atopic diseases in infancy. The German multicenter atopy study (MAS-90). Pediatr Allergy Immunol. 1994;5(S5):19–25.
100. Dex S, Joshi H. Children of the 21st century: from birth to nine months. Policy Press. 2005.
101. Clarisse B et al. The Paris prospective birth cohort study: which design and who participates? Eur J Epidemiol. 2007;22(3):203–10.
102. Ege MJ et al. Prenatal exposure to a farm environment modifies atopic sensitization at birth. J Allergy Clin Immunol. 2008;122(2):407–12–412.e1-4.
103. Wijga A et al. Are children at high familial risk of developing allergy born into a low risk environment? The PIAMA birth cohort study. Clin Exp Allergy. 2001;31(4):576–81.
104. Hagendorens MM et al. Perinatal risk factors for sensitization, atopic dermatitis and wheezing during the first year of life (PIPO study). Clin Exp Allergy J Br Soc Allergy Clin Immunol. 2005;35(6):733–40.
105. Havstad S et al. Effect of prenatal indoor pet exposure on the trajectory of total IgE levels in early childhood. J Allergy Clin Immunol. 2011;128(4):880–885.e4.