Modelling Wheezing Spells Identifies Phenotypes with Different Outcomes and Genetic Associates

Sadia Haider[1], Raquel Granell[2], John Curtin[3], Sara Fontanella[1], Alex Cucco MSc[1], Stephen Turner[4,5], Angela Simpson[3], Graham Roberts[6,7,8], Clare S Murray[3], John W. Holloway[6,7], Graham Devereux[9], Paul Cullinan[1], Syed Hasan Arshad[7,8,10], Adnan Custovic[1]

[1]National Heart and Lung Institute, Imperial College London, UK

[2]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, UK

[3]Division of Infection, Immunity and Respiratory Medicine, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, UK

[4]Royal Aberdeen Children's Hospital NHS Grampian Aberdeen, AB25 2ZG, UK

[5]Child Health, University of Aberdeen, Aberdeen, UK

[6]Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK

[7] NIHR Southampton Biomedical Research Centre, University Hospitals Southampton NHS Foundation Trust, Southampton, UK

[8]David Hide Asthma and Allergy Research Centre, Isle of Wight, UK

[9]Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK

[10]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK

Correspondence to: Adnan Custovic MD PhD FMedSci, Imperial College London, a.custovic@imperial.ac.uk

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Contributions: SH, AC, SF and RG conceived and planned the study, and wrote the manuscript. SH, RG, SF and JC analysed the data. All authors contributed to the interpretation of the results. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Funding: Consortium is funded by the UK Medical Research Council (MRC) Programme Grant MRCMR/S025340/1 and was funded through MRC grants G0601361 and MR/K002449/1. RG is in part funded through Wellcome Trust Strategic Award 108818/15/Z. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. MAAS was supported by the Asthma UK Grants No 301 (1995-1998), No 362 (1998-2001), No 01/012 (2001-2004), No 04/014 (2004-2007), BMA James Trust (2005) and the JP Moulton Charitable Foundation (2004-2016), The North west Lung Centre Charity (1997-current) and the Medical Research Council (MRC) grant MR/L012693/1 (2014-2018)

Abstract word count: 250

Word count: 3496

At a Glance

What is the current scientific knowledge on this subject?

Longitudinal modelling of current wheezing identified similar phenotypes, but their characteristics often differ between studies.

What does this study add to the field?

Transformation of binary wheeze data into a set of multi-dimensional variables better captures the temporal characteristics of wheeze development and provides a more robust input for phenotype derivation. Modelling using multi-dimensional variables of wheezing spells identified a stable and consistent architecture of wheezing illness, including a novel intermittent phenotype associated with early lung function decline to early adulthood. Different wheezing phenotypes are underpinned by unique mechanisms and genetic associates.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

This article has an online data supplement, which is accessible from this issue's table of content online at www.atsjournals.org.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

ABSTRACT

Background: Longitudinal modelling of current wheezing identified similar phenotypes, but their characteristics often differ between studies. We propose that more comprehensive description of wheeze may better describe trajectories than binary information on presence/absence of wheezing.

Methods: We derived 6 multi-dimensional variables of wheezing spells from birth to adolescence (including duration, temporal sequencing, and the extent of persistence/recurrence). We applied Partition-Around-Medoids clustering on these variables to derive phenotypes in five birth cohorts. We investigated within- and between-phenotype differences compared to binary latent class analysis models (LCA-phenotypes), and ascertained associations of these phenotypes with asthma and lung function, and with polymorphisms in asthma loci 17q12-21 and *CDHR3*.

Findings: Analysis among 7719 participants with complete data identified 5 spell-based wheeze phenotypes with high degree of certainty: Never (NWZ-54.1%), Early-transient (ETW-23.7%), Late-onset (LOW-6.9%), Persistent (PEW-8.3%), and a novel phenotype, Intermittent wheeze (INT-6.9%). FEV_1/FVC was lower in PEW and INT compared to ETW and LOW, and declined from age 8 years to adulthood in INT. 17q12-21 and *CDHR3* polymorphisms were associated with higher odds of PEW and INT, but not ETW or LOW. LCA- and spell-based-phenotypes appeared similar, but within-phenotype individual trajectories and phenotype allocation differed substantially. The spell-based approach was much more robust in dealing with missing data, and the derived clusters more stable and internally homogenous.

Conclusions: Modelling of spell variables identified a novel intermittent wheeze phenotype associated with lung function decline to early adulthood. Using multi-dimensional spells variables may better capture wheeze development and provide a more robust input for phenotype derivation.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

INTRODUCTION

Wheeze in most children remits by school age, but in others may persist, with or without periods of remission. Over the past decades, a substantial effort has been devoted to understanding the heterogeneity of childhood wheezing illness, using both hypothesis-driven approaches (in which phenotypes are specified *a priori* based on clinical insights (1)) and data-driven ones, which incorporate a variety of multivariate statistical and machine learning methodologies (2). The latter have largely used latent class modelling, such as latent class analysis (LCA), in which repeated information of wheeze presence is used to uncover the temporal patterns over a specified time interval (3-15). These different symptom patterns may indicate distinct causes and biological mechanisms (16, 17), and their discovery may facilitate stratified treatment (18). However, to facilitate identification of genetic associates and underlying mechanisms, phenotypes should be internally homogenous and consistent between different populations/studies.

The number of phenotypes reported in previous analyses which used LCA varied by study, but four were identified in all cohorts (19): Never/infrequent wheeze, Transient early, Late−onset and Persistent wheeze. Some analyses identified one or two further “intermediate” phenotypes (3, 4, 20), which mostly arose from transient or late-onset patterns (21). However, although phenotypes in different studies are usually designated with the same name, they often differ in temporal trajectories, distributions within a population, and associated risk factors (19, 22). These differences are in part a consequence of the sample size and the timing and frequency of data collection (21). Furthermore, the confidence with which individuals are assigned to a phenotype varies across phenotypes, and a substantial number of children in such analyses are classified imprecisely (e.g. individuals with identical wheeze patterns may be assigned to different phenotypes, or, individual trajectories may not follow wheeze patterns suggested by the phenotype label (13, 21, 23)).

We propose that within−class heterogeneity and inaccurate allocation of individual children may, in part, be responsible for a lack of consistent associations of discovered phenotypes with risk factors

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

(24), and may adversely impact the ability to identify phenotype-specific genetic associates and underlying mechanisms. We hypothesise that incorporating more comprehensive description of wheeze may better describe wheeze trajectories and derive more within-phenotype homogeneity to facilitate better understanding of their differing aetiology. To address our hypothesis, we drew on research in other fields, specifically the 'spells' approach pioneered in the social sciences research on poverty dynamics (25-28), to move from the point prevalence of current wheeze, to a dynamic approach which takes into account the duration of wheezing spells, their temporal sequencing, and the extent of persistence and recurrence (further details in the Supplementary Introduction). To this end, we first developed a set of multi-dimensional variables to describe more comprehensively the temporal variation of wheeze, and then applied a clustering approach based on the Partition Around Medoids (PAM) algorithm (29) on these variables. We then investigated variation within and between phenotypes from binary (LCA) and indicator-based (PAM) models to ascertain whether we achieved increased within-phenotype homogeneity, and investigated the associations of the derived clusters with early−life factors and asthma−related outcomes in adolescence. Finally, we tested the hypothesis that phenotypes defined using this approach have distinct genetic associates by investigating their associations with the known asthma loci (*17q12-21* and *CDHR3*).

METHODS

Study design, setting and participants

The Study Team for Early Life Asthma Research (STELAR) consortium (30) brings together five UK population−based birth cohorts: Avon Longitudinal Study of Parents and Children (ALSPAC) (31), Ashford (32), Isle of Wight (IOW) (33) and Aberdeen (SEATON) (34) cohorts, and Manchester Asthma and Allergy Study (MAAS) (35). The cohorts are described in detail in the Supplementary Appendix. All studies were approved by research ethics committees. Informed consent was obtained from parents, and study participants gave their assent/consent when applicable. Data were harmonised into the web−based knowledge management platform to enable joint analyses (30).

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Data sources and definition of variables

Validated questionnaires were completed on multiple occasions from infancy to adolescence (23). The cohort−specific time points and sample sizes are shown in Table S1. For the analyses of pooled data, we defined epochs based on the data availability at shared time points across cohorts: infancy (½-1 year); early childhood (2-3 years); pre-school/early school (4-5 years); middle childhood (8-10 years); and adolescence (14-18 years) (23). For each child, we derived 6 wheeze variables:

1 Age of the first episode

2 Age of the last recorded episode

3 Total number of separate records over the observation period

4 Duration of the longest spell based on the number of consecutive records of wheeze

5 Total number of separate wheeze spells

6 Spell type: a categorical variable defined as 0=no wheeze, 1=single spell, 2=intermittent spells (defined as at least two non-consecutive spells of wheeze of any length).

An illustrative example of the derivation of the variables is shown in Table S2.

We performed spirometry in adolescence in all cohorts, and in ALSPAC and MAAS on at least three follow-ups from school-age to early adulthood. We recorded FEV_1 and FVC and expressed data as z-scores for each population.

Skin testing was carried out in early/mid-school age in all cohorts, and on six follow-ups in MAAS. Definition of all variables can be found in the Supplementary appendix.

Statistical analysis

We analysed pooled data from participating cohorts. Figure S1 provides an overview of the analytical steps. A detailed description is provided in the Supplementary appendix.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Wheeze phenotypes from infancy to adolescence from 6 derived variables: To derive longitudinal wheeze patterns captured by the multi-dimensional variables, we used PAM (29) algorithm coupled with the Wishart distance for mixed data (36), initially among 7719 participants with complete data on wheezing at all five time-points. To investigate whether our findings were influenced by missing data, we adopted the framework of Basagaña *et al*. (37) which integrates multiple imputation (38) into cluster analysis, and applied it to data of 15,848 participants with at least 2 observations.

Comparison of wheeze phenotypes derived using binary LCA and spells PAM approaches: We first repeated analyses from our previous study which used LCA to identify 5 wheeze phenotypes in the same 7719 participants (Never/infrequent, Pre−school remitting, Mid−childhood remitting, Persistent and Late−onset (23)), and assigned participants to phenotypes according to the maximum posterior probability. We then compared the within-class homogeneity of both models. We checked the stability of cluster allocations using the adjusted Rand index (ARI) (39), and plotted the magnitude of transitions of phenotype membership between models using alluvial plots.

Association of spell-based PAM-phenotypes with early−life risk factors and clinical outcomes in adolescence: We used multinomial logistic regression to ascertain early−life associates of each PAM-phenotype and examine their relationship with doctor−diagnosed asthma and asthma medication use in adolescence; results are reported as relative risk ratios (RRR) with 95% confidence intervals (CI). Associations with lung function (Z−scores for FEV_1, FVC and FEV_1/FVC adjusted for height, age and sex) were investigated using linear regression. Models were adjusted for potential confounders, including maternal history of asthma, maternal smoking and low birth weight.

Genetic associates of spell-based PAM-phenotypes: We investigated the association of derived clusters with 17q12-21 SNPs (Table S3) and *CDHR3* SNP rs6967330 (40). We selected one representative 17q12-21 SNP per linkage disequilibrium block, leaving rs7216389, rs4795408 and rs3894194 in the final analysis. We tested additive model using multinomial logistic regression.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

RESULTS

Characteristics of the study population

Of 7719 children with complete data on wheezing, 50.4% were male. At the follow-up in adolescence, 12.9% had current asthma and 11.4% reported using asthma medication. Demographic characteristics are shown in Table S4, and wheeze prevalence in Table S5. The prevalence of current wheeze decreased from 22.8% in infancy to 13.7% in adolescence.

Wheeze phenotypes obtained using 6 derived variables and PAM algorithm

A five−cluster solution was selected as the optimal based on statistical fit (Figure S2). After inspection of trajectories for each cluster (Figure 1), the clusters (phenotypes) were characterised as: (1) Never wheeze (NWZ) (54.1%); (2) Early transient (ETW) (23.7%); (3) Late-onset (LOW) (6.9%); (4) Persistent (PEW) (8.3%); and (5) Intermittent (INT) wheeze (6.9%). The same 5-class structure was evident when we modelled each cohort separately (Table S6), and the optimal solution was stable to changes in sample size (Table S7).

Impact of missing data on cluster derivation: Detailed analysis is shown in Supplementary results. The optimal solution from the model using 15,848 individuals with $\geq$2 observations was very similar to that from 7719 participants with complete data (Table S8). Children were assigned to clusters with a high degree of certainty (Table S9). There was a very high agreement between phenotype assignment of individual children when using complete or imputed data (ARI=0.94); only 195/7719 (2.5%) children changed phenotype allocation (Figure S3).

Comparison of wheezing phenotypes derived using binary LCA and spells PAM approaches

Figures S4 shows latent classes (phenotypes) identified by LCA. Phenotypes derived using the two methods among the same 7719 participants appeared very similar, and four appeared identical (NWZ, ETW, PEW and LOW) (Figures 1 and S4). However, the within-phenotype structure differed substantially (Figure 2). For example, in PAM-NWZ cluster no participants reported wheezing at any

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

time point (Figure 2a), while in LCA-NWZ 10% reported occasional wheezing (Figure 2c). In PAM-ETW, no participants reported wheezing after age 10 years, and nobody in PAM-LOW wheezed before age 8; in contrast, in the LCA-ETW class, 8% reported wheeze up to age 18 years, and wheeze before age 10 was present among 42% in LCA-LOW.

Figure 2 (b/d) and Table S10 show the distribution of wheeze variables between phenotypes from the two approaches. In PAM-LOW, the earliest observed age of wheeze onset was 7 years later than in LCA-LOW. PAM-PEW only contained children with a long single spell of wheeze, whereas subjects in the LCA-PEW also had intermittent spells.

We further investigated the differences between individual allocations to PAM and LCA phenotypes for all 32 possible wheeze sequences across the 5 time points (Table S11). We did not observe any inconsistencies across cohorts in the PAM model (i.e., the same sequences were always assigned to the same cluster). In contrast, children with identical sequences were assigned by LCA to different phenotypes (e.g., “0-1-0-1-0” was assigned to 3 different LCA phenotypes, while PAM spell-based analysis always assigned this sequence to the INT phenotype).

Figure S5 shows differences in individual assignment to PAM and LCA phenotypes. One quarter of subjects transitioned to a different phenotype. A higher stability was observed for ETW and LOW (>70%), but was relatively poor in the PEW (60%). Children in PAM-INT cluster transitioned from all LCA phenotypes.

Finally, we applied PAM algorithm to the binary current wheeze variable (yes/no) to investigate whether the algorithm or the transformation to spell-based variables gave rise to homogeneous phenotypes. A 5-cluster solution was optimal; however, the clusters resembled LCA phenotypes (with no INT wheeze) and were structurally internally much more heterogeneous than phenotypes obtained using the derived variables (Figure S6). Therefore, it is likely that the derived variables were, primarily, the precursor for deriving more homogeneous phenotypes.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Association of spell-based phenotypes with early–life risk factors and asthma-related outcomes

Family history, early–life factors and environmental exposures: Univariable analyses are shown in Table S12. Table S13 shows results of multivariable logistic regression models. Males had higher risk of developing PEW, ETW, and INT, but not LOW. Maternal asthma and parental smoking were associated with all four clusters. Low birth weight was associated with ETW, INT and PEW (with the strongest association with PEW), but not with LOW.

Asthma: Compared with NWZ, all 4 wheeze clusters were associated with a higher risk of asthma diagnosis and medication use in adolescence (Table 1). The associations were strongest for PEW and weakest for ETW (e.g., the risk of using asthma medication was approximately 14-fold higher for PEW than ETW). Variability in the proportion of asthmatics by spell-based phenotype and the proportion of subjects with asthma diagnosis in adolescence in each phenotype are shown in Figure 3; of note, 5.7% of children with asthma diagnosis in adolescence never reported wheezing.

Allergic sensitization: All phenotypes were associated with sensitisation in early-school age (Table S13), with the magnitude of risk being higher for PEW and LOW. Trajectories of sensitization from infancy to adolescence in MAAS were almost identical in PEW, INT and LOW, and differed from those in NWZ and ETW (Figure 4), i.e., highly concordant longitudinal sensitization patterns were associated with different wheeze phenotypes. In general, wheeze preceded sensitization in PEW and INT clusters, while sensitization preceded wheeze in LOW.

Lung function: FEV_1/FVC in adolescence was lower in all four wheeze phenotypes compared to children who never wheezed, with those in PEW having the lowest lung function, markedly lower compared to NWZ (z-score: -0.71; 85% CI [-0.83, -0.59], P<0.0001) (Table 1). FVC was similar across clusters. Longitudinal lung function was available in 6729, 4567 and 3749 participants at ages 8, 15 and 24 years respectively in ALSPAC, and 790, 801, 630 and 504 at ages 8, 11, 16 and 20 in MAAS. FEV_1/FVC was significantly lower in all wheeze phenotypes compared to NWZ throughout the follow

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

up (Figure 5), and was consistently lower in PEW and INT compared to ETW and LOW (Table S14). FEV_1/FVC declined from age 8 years to early adulthood in INT, but not other phenotypes.

Association between spell-based phenotypes and genetic variants in *17q12-21* and *CDHR3*

9655 subjects of white European ancestry had genotyping data and were included in the meta-analysis of genetic associations. Figure 6 shows forest plots of the associations for representative SNPs. Sub-group level p-values are presented in Table S15. We found strong evidence of an association between all 17q12-21 SNPs and PEW. INT was also associated with 17q12-21 SNPs. However, we found little evidence of an association between 17q12-21 SNPs and ETW and LOW.

We found strong evidence of an association between *CDHR3* SNP rs6967330 and PEW (OR 1.45, 95% CI 1.03-2.04) and INT (1.40, 1.04-1.89), but there was no association with ETW and LOW clusters.

DISCUSSION

We applied a framework which focussed on wheezing spells to describe the temporal patterns of wheeze from infancy to adolescence. Our results suggest that this approach better captures wheeze development than presence/absence of wheezing alone and provides a more robust input for data-driven phenotype derivation. It is much more robust in dealing with missing data, and the derived clusters are stable and internally homogenous. Our spells-based analysis applied to data from five population-based birth cohorts identified a novel wheezing phenotype, intermittent wheeze, to which ~7% of participants were assigned. FEV_1/FVC trajectory from school-age to physiological peak in early adulthood showed consistently diminished lung function in all four wheeze phenotypes determined using the spells-based approach compared to never wheezers, and in Persistent and Intermittent compared to Transient-early and Late-onset wheezing. Lung function declined from age 8 years to early adulthood in intermittent, but not other phenotypes. Finally, associations with 17q12-21 and *CDHR3* SNPs differed across wheezing phenotypes, and carriers of risk variants had significantly increased risk for persistent and intermittent, but not of transient or late-onset wheeze.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Wheezing phenotypes developed using spells appeared more clinically intuitive than those derived based on wheeze presence/absence. For example, no subjects in spell-based-ETW reported wheezing after age 10 years, and nobody in LOW wheezed before age 10 years; in contrast, in the LCA-ETW, some children reported wheeze to age 18 years, and early-life wheeze was reported in some individuals assigned to LCA-LOW. In spell-based-LOW, the earliest observed age of wheeze onset was 7 years later than in LCA-LOW.

Within-class heterogeneity may dilute associations with biomarkers, genetic variants, and environmental factors. Therefore, for such analyses, phenotypes derived using data-driven methods should be homogenous, and individual patterns of symptoms within each phenotype should be distinct from individuals in other subgroups. Our previous LCA showed that a substantial number of children are classified imprecisely using binary information on wheeze, particularly when an individual's posterior probability of assignment is <0.80 (21). Similarly, a recent US study which derived wheeze phenotypes using LCA found that one third of subjects had a posterior probability <0.80 (13). Our current analysis demonstrates that when using binary representation of wheeze, some wheeze patterns are not assigned to phenotypes with high precision, and consequently, individuals with the same longitudinal wheezing patterns can be assigned to different phenotypes. The intermittent patterns contributed to substantial within-class heterogeneity when using binary data in both LCA and PAM models. Once the spells approach isolated these intermittent patterns, ETW, LOW and PEW were more internally homogeneous, and a novel INT cluster emerged.

Our previous analysis in the same study population showed that data imputation has a major impact on the assignment of individual participants to different phenotypes in LCA (e.g., ~40% of children switched from Early−onset middle−childhood remitting to PEW from model with complete data to that with imputed data (23)). In contrast, in the current study, there was a remarkably high agreement between assignment of individuals into clusters when using complete or imputed data,

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

and only 2.5% of children changed phenotype. This is of key importance for longitudinal studies in which data missingness is inevitable, and for genetic analyses in which large sample size is essential.

The important question as to whether different longitudinal wheezing phenotypes are underpinned by unique pathophysiological mechanisms has been asked by Koppelman and Kersten (41) in an editorial following the recent finding from the CREW consortium which investigated the association of 17q12-21 SNPs with LCA-derived phenotypes (13). In this study, contrary to the hypothesis of differential genetic associations of different wheeze phenotypes, associations between multiple 17q12-21 SNPs were similar for all LCA phenotypes, suggesting that all wheezing phenotypes have shared genetic origin in relation to this locus (13). In contrast, we found a clear differential association of genetic markers between phenotypes derived using spells-based variables. We found no association of the SNPs in this locus with transient and late-onset wheezing, and our results do not support the notion that the 17q locus should be considered a "wheezing locus".

Both 17q21 locus and *CDHR3* are linked to differential susceptibility to infection by rhinoviruses (42, 43), and our data suggest that such susceptibility is common and important for early-onset non-transient phenotypes (both persistent and intermittent). However, most children who wheeze in early life stop wheezing by school-age (~2/3 in our data set, all of whom clustered to spell-based-ETW), and known genetic markers of susceptibility to rhinoviruses were not apparent in this group. This is consistent with recent data showing that even among children with severe recurrent preschool wheeze, ~50% had no evidence of either inflammation or infection in their lower airways (44). It is possible that diminished lung function in early childhood (as suggested by the seminal study from Tucson cohort (45) and indirectly confirmed in one of our cohorts (46, 47)), is associated with poor growth in early childhood (48) or specific genetic susceptibility (49, 50), and is a principal cause of early-onset transient wheezing, while susceptibility to viruses may contribute to persistence and exacerbations. We cannot exclude that the immune response to other viruses (such as RSV) may also be important in ETW (51). Our data also suggest that LOW (which in the current analysis started

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

after age 10 years) in most children may not be associated with susceptibility to viruses but is predominantly allergic airway disease (as suggested by the analysis of the pattern of *in vitro* immune responses to viruses (52)). In these individuals, allergen exposure may be the principal contributor to severity and exacerbations (53). However, it is important to emphasize that all wheeze phenotypes were associated with diminished lung function in adolescence and early adulthood, with the greatest impairment in PEW and INT. This is a precursor of COPD (54-56), early all-cause mortality (57) and early-onset cardio−vascular, respiratory and metabolic comorbidities (58).

We found that 5.7% of children with asthma diagnosis in adolescence belonged to the NWZ (and a similar proportion to the ETW group). This emphasises the heterogeneity of doctor-diagnosed asthma at the population level, and the fact that children with other respiratory symptoms such as cough (even in the absence of wheezing) are diagnosed as being asthmatic.

One limitation of our study is that the population is not ethnically diverse. In addition, early life pulmonary/airway function tests were not performed, which limits the inference to the potential role of pre-morbid lung function. We also acknowledge that our study was not able to investigate the relationship between wheeze treatment, disease severity, and patterns of wheeze spells. With respect to genetic analyses, further investigations are needed at a genome-wide level to help distinguish mechanisms of early-life wheeze and subsequent asthma.

In conclusion, our data are consistent with the notion that in addition to shared pathophysiology, distinct wheezing phenotypes are underpinned by unique mechanisms and genetic associates. Modelling using multi-dimensional variables of wheezing spells identified a stable and consistent architecture of wheezing illness, including a novel intermittent phenotype associated with early lung function decline to early adulthood. We suggest that the transformation of binary data into a set of multi-dimensional variables may better capture the temporal characteristics of wheeze development and may provide a more robust input for phenotype derivation.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

This article is dedicated to the memory of our wonderful colleague and friend Professor John Henderson (1958-2019), whose contribution to the understanding of the heterogeneity of childhood asthma cannot be overstated. Rainbow-chasers and UNICORN riders forever.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

REFERENCES

1. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ, Bean J, Bianchi H, Curtiss J, Ey J, Sanguineti A, Smith B, Vondrak T, West N, Mclellan M. Asthma and Wheezing in the First 6 Years of Life. *New Engl J Med* 1995; 332: 133-138.

2. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. *Current allergy and asthma reports* 2015; 15: 38.

3. Henderson J, Granell R, Heron J, Sherriff A, Simpson A, Woodcock A, Strachan DP, Shaheen SO, Sterne JAC. Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood. *Thorax* 2008; 63: 974-980.

4. Granell R, Henderson AJ, Sterne JA. Associations of wheezing phenotypes with late asthma outcomes in the Avon Longitudinal Study of Parents and Children: A population-based birth cohort. *J Allergy Clin Immun* 2016; 138: 1060-1070.

5. Granell R, Sterne J, Savenije O, Kerkhof M, Smit H, Jongste JC, Postma DS, Koppelman G, Henderson J. Identification And Replication Of Wheezing Phenotypes Using Longitudinal Latent Class Analysis. American Thoracic Society 2010 International Conference. New Orleans 2010. p. A6242.

6. Fitzpatrick AM, Bacharier LB, Guilbert TW, Jackson DJ, Szefler SJ, Beigelman A, Cabana MD, Covar R, Holguin F, Lemanske RF, Jr., Martinez FD, Morgan W, Phipatanakul W, Pongracic JA, Zeiger RS, Mauger DT, AsthmaNet NN. Phenotypes of Recurrent Wheezing in Preschool Children: Identification by Latent Class Analysis and Utility in Prediction of Future Exacerbation. *J Allergy Clin Immunol Pract* 2019; 7: 915-924 e917.

7. Depner M, Fuchs O, Genuneit J, Karvonen AM, Hyvarinen A, Kaulek V, Roduit C, Weber J, Schaub B, Lauener R, Kabesch M, Pfefferle PI, Frey U, Pekkanen J, Dalphin JC, Riedler J, Braun-Fahrlander

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

C, von Mutius E, Ege MJ, Group PS. Clinical and epidemiologic phenotypes of childhood asthma. *Am J Respir Crit Care Med* 2014; 189: 129-138.

8. Spycher BD, Silverman M, Pescatore AM, Beardsmore CS, Kuehni CE. Comparison of phenotypes of childhood wheeze and cough in 2 independent cohorts. *J Allergy Clin Immunol* 2013; 132: 1058-1067.

9. Belgrave DCM, Simpson A, Semic-Jusufagic A, Murray CS, Buchan I, Pickles A, Custovic A. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. *J Allergy Clin Immunol* 2013; 132: 575-583 e512.

10. 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: 974-981.

11. Sordillo JE, Coull BA, Rifas-Shiman SL, Wu AC, Lutz SM, Hivert MF, Oken E, Gold DR. Characterization of longitudinal wheeze phenotypes from infancy to adolescence in Project Viva, a prebirth cohort study. *J Allergy Clin Immunol* 2020; 145: 716-719 e718.

12. Kotecha SJ, Watkins WJ, Lowe J, Granell R, Henderson AJ, Kotecha S. Comparison of the Associations of Early-Life Factors on Wheezing Phenotypes in Preterm-Born Children and Term-Born Children. *Am J Epidemiol* 2019; 188: 527-536.

13. Hallmark B, Wegienka G, Havstad S, Billheimer D, Ownby D, Mendonca EA, Gress L, Stern DA, Myers JB, Khurana Hershey GK, Hoepner L, Miller RL, Lemanske RF, Jackson DJ, Gold DR, O'Connor GT, Nicolae DL, Gern JE, Ober C, Wright AL, Martinez FD, Echo C. Chromosome 17q12-21 Variants are Associated with Multiple Wheezing Phenotypes in Childhood. *Am J Respir Crit Care Med* 2021;203(7):864-870.

14. Odling M, Wang G, Andersson N, Hallberg J, Janson C, Bergstrom A, Melen E, Kull I. Characterization of asthma trajectories from infancy to young adulthood. *J Allergy Clin Immunol Pract* 2021;9(6):2368-2376.e3.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

15. Savenije OE, Granell R, Caudri D, Koppelman GH, Smit HA, Wijga A, de Jongste JC, Brunekreef B, Sterne JA, Postma DS. Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA. *J Allergy Clin Immunol* 2011; 127: 1505-1512. e1514.

16. Saria S, Goldenberg A. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems* 2015; 30: 70-75.

17. 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: 121-123.

18. Saglani S, Custovic A. Childhood Asthma: Advances Using Machine Learning and Mechanistic Studies. *Am J Respir Crit Care Med* 2019; 199: 414-422.

19. Owora AH, Zhang Y. Childhood wheeze trajectory-specific risk factors: A systematic review and meta-analysis. *Pediatr Allergy Immunol* 2021; 32: 34-50.

20. Granell R, Sterne JA, Savenije O, Kerkhof M, Smit HA, Jongste JD, Postma D, Koppelman G, Henderson J. Identification and Replication of Wheezing Phenotypes using Longitudinal Latent Class Analysis. *Am J Resp Crit Care* 2010; 181.

21. Oksel C, Granell R, Mahmoud O, Custovic A, Henderson AJ, Stelar, Breathing Together i. Causes of variability in latent phenotypes of childhood wheeze. *J Allergy Clin Immunol* 2019; 143: 1783-1790 e1711.

22. 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.

23. Oksel C, Granell R, Haider S, Fontanella S, Simpson A, Turner S, Devereux G, Arshad SH, Murray CS, Roberts G, Holloway JW, Cullinan P, Henderson J, Custovic A, Stelar investigators bTi. Distinguishing Wheezing Phenotypes from Infancy to Adolescence. A Pooled Analysis of Five Birth Cohorts. *Ann Am Thorac Soc* 2019; 16: 868-876.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

24. 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: 921-936.

25. Bane MJ, Ellwood DT. Slipping into and out of Poverty - the Dynamics of Spells. *J Hum Resour* 1986; 21: 2-23.

26. Layte R, Whelan CT. Moving in and out of poverty - The impact of welfare regimes on poverty dynamics in the EU. *Eur Soc* 2003; 5: 167-191.

27. Mendola D, Busetta A. The Importance of Consecutive Spells of Poverty: A Path-Dependent Index of Longitudinal Poverty. *Rev Income Wealth* 2012; 58: 355-374.

28. Stevens AH. Climbing out of poverty, falling back in - Measuring the persistence of poverty over multiple spells. *J Hum Resour* 1999; 34: 557-588.

29. Partitioning Around Medoids (Program PAM). *Finding Groups in Data* 1990.

30. Custovic A, Ainsworth J, Arshad H, Bishop C, Buchan I, Cullinan P, Devereux G, Henderson J, Holloway J, Roberts G, Turner S, Woodcock A, Simpson A. 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.

31. Golding J, Pembrey M, Jones R. ALSPAC-the Avon longitudinal study of parents and children. I. Study methodology. *Paediatric and perinatal epidemiology* 2001; 15: 74-87.

32. Cullinan P, MacNeill SJ, Harris JM, Moffat S, White C, Mills P, Newman Taylor AJ. Early allergen exposure, skin prick responses, and atopic wheeze at age 5 in English children: a cohort study. *Thorax* 2004; 59: 855-861.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

33. Arshad SH, Holloway JW, Karmaus W, Zhang H, Ewart S, Mansfield L, Matthews S, Hodgekiss C, Roberts G, Kurukulaaratchy R. Cohort Profile: The Isle Of Wight Whole Population Birth Cohort (IOWBC). *Int J Epidemiol* 2018; 47: 1043-1044i.

34. Martindale S, McNeill G, Devereux G, Campbell D, Russell G, Seaton A. Antioxidant intake in pregnancy in relation to wheeze and eczema in the first two years of life. *Am J Resp Crit Care* 2005; 171: 121-128.

35. Custovic A, Simpson BM, Murray CS, Lowe L, Woodcock A, Asthma NACM, Allergy Study G. The National Asthma Campaign Manchester Asthma and Allergy Study. *Pediatr Allergy Immunol* 2002; 13: 32-37.

36. Wishart D. k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values. In: Schwaiger M, Opitz O, editors. Exploratory Data Analysis in Empirical Research. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 216-226.

37. Basagana X, Barrera-Gomez J, Benet M, Anto JM, Garcia-Aymerich J. A Framework for Multiple Imputation in Cluster Analysis. *American Journal of Epidemiology* 2013; 177: 718-725.

38. Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons Inc; 2Rev Ed edition (24 Sept. 2002); 2002.

39. Hubert L, Arabie P. Comparing partitions. *Journal of Classification* 1985; 2: 193-218.

40. Bonnelykke K, Sleiman P, Nielsen K, Kreiner-Moller E, Mercader JM, Belgrave D, den Dekker HT, Husby A, Sevelsted A, Faura-Tellez G, Mortensen LJ, Paternoster L, Flaaten R, Molgaard A, Smart DE, Thomsen PF, Rasmussen MA, Bonas-Guarch S, Holst C, Nohr EA, Yadav R, March ME, Blicher T, Lackie PM, Jaddoe VW, Simpson A, Holloway JW, Duijts L, Custovic A, Davies DE, Torrents D, Gupta R, Hollegaard MV, Hougaard DM, Hakonarson H, Bisgaard H. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet* 2014; 46: 51-55.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

41. Koppelman GH, Kersten ETG. Understanding How Asthma Starts: Longitudinal Patterns of Wheeze and the Chromosome 17q Locus. *Am J Respir Crit Care Med* 2021; 203: 793-795.

42. Zhang Y, Willis-Owen SAG, Spiegel S, Lloyd CM, Moffatt MF, Cookson W. The ORMDL3 Asthma Gene Regulates ICAM1 and Has Multiple Effects on Cellular Inflammation. *Am J Respir Crit Care Med* 2019; 199: 478-488.

43. Basnet S, Bochkov YA, Brockman-Schneider RA, Kuipers I, Aesif SW, Jackson DJ, Lemanske RF, Jr., Ober C, Palmenberg AC, Gern JE. CDHR3 Asthma-Risk Genotype Affects Susceptibility of Airway Epithelium to Rhinovirus C Infections. *Am J Respir Cell Mol Biol* 2019; 61: 450-458.

44. Robinson PFM, Fontanella S, Ananth S, Martin Alonso A, Cook J, Kaya-de Vries D, Polo Silveira L, Gregory L, Lloyd C, Fleming L, Bush A, Custovic A, Saglani S. Recurrent Severe Preschool Wheeze: From Prespecified Diagnostic Labels to Underlying Endotypes. *Am J Respir Crit Care Med* 2021; 204: 523-535.

45. Morgan WJ, Stern DA, Sherrill DL, Guerra S, Holberg CJ, Guilbert TW, Taussig LM, Wright AL, Martinez FD. Outcome of asthma and wheezing in the first 6 years of life: follow-up through adolescence. *Am J Respir Crit Care Med* 2005; 172: 1253-1258.

46. Lowe LA, Simpson A, Woodcock A, Morris J, Murray CS, Custovic A, Asthma NACM, Allergy Study G. Wheeze phenotypes and lung function in preschool children. *Am J Respir Crit Care Med* 2005; 171: 231-237.

47. Belgrave DC, Buchan I, Bishop C, Lowe L, Simpson A, Custovic A. Trajectories of lung function during childhood. *Am J Respir Crit Care Med* 2014; 189: 1101-1109.

48. Voraphani N, Stern DA, Zhai J, Wright AL, Halonen M, Sherrill DL, Hallberg J, Kull I, Bergström A, Murray CS, Lowe L, Custovic A, Morgan W, Martinez FD, Melén E, Simpson A, Guerra S. Early Origins of Spirometric Restriction: The Role of Growth and Nutrition. *Lancet Respiratory Medicine* 2022;10(1):59-71.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

49. Simpson A, Custovic A, Tepper R, Graves P, Stern DA, Jones M, Hankinson J, Curtin JA, Wu J, Blekic M, Bukvic BK, Aberle N, Marinho S, Belgrave D, Morgan WJ, Martinez FD. Genetic variation in vascular endothelial growth factor-a and lung function. *Am J Respir Crit Care Med* 2012; 185: 1197-1204.

50. Simpson A, Maniatis N, Jury F, Cakebread JA, Lowe LA, Holgate ST, Woodcock A, Ollier WE, Collins A, Custovic A, Holloway JW, John SL. Polymorphisms in a disintegrin and metalloprotease 33 (ADAM33) predict impaired early-life lung function. *Am J Respir Crit Care Med* 2005; 172: 55-60.

51. Raita Y, Perez-Losada M, Freishtat RJ, Harmon B, Mansbach JM, Piedra PA, Zhu Z, Camargo CA, Hasegawa K. Integrated omics endotyping of infants with respiratory syncytial virus bronchiolitis and risk of childhood asthma. *Nat Commun* 2021; 12: 3601.

52. Custovic A, Belgrave D, Lin L, Bakhsoliani E, Telcian AG, Solari R, Murray CS, Walton RP, Curtin J, Edwards MR, Simpson A, Rattray M, Johnston SL. Cytokine Responses to Rhinovirus and Development of Asthma, Allergic Sensitization, and Respiratory Infections during Childhood. *Am J Respir Crit Care Med* 2018; 197: 1265-1274.

53. Custovic A, Custovic D, Kljaic Bukvic B, Fontanella S, Haider S. Atopic phenotypes and their implication in the atopic march. *Expert review of clinical immunology* 2020; 16: 873-881.

54. Bui DS, Lodge CJ, Burgess JA, Lowe AJ, Perret J, Bui MQ, Bowatte G, Gurrin L, Johns DP, Thompson BR, Hamilton GS, Frith PA, James AL, Thomas PS, Jarvis D, Svanes C, Russell M, Morrison SC, Feather I, Allen KJ, Wood-Baker R, Hopper J, Giles GG, Abramson MJ, Walters EH, Matheson MC, Dharmage SC. 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: 535-544.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

55. Belgrave DCM, Granell R, Turner SW, Curtin JA, Buchan IE, Le Souef PN, Simpson A, Henderson AJ, Custovic A. 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: 526-534.

56. Berry CE, Billheimer D, Jenkins IC, Lu ZJ, Stern DA, Gerald LB, Carr TF, Guerra S, Morgan WJ, Wright AL, Martinez FD. A Distinct Low Lung Function Trajectory from Childhood to the Fourth Decade of Life. *Am J Respir Crit Care Med* 2016; 194: 607-612.

57. Vasquez MM, Zhou M, Hu C, Martinez FD, Guerra S. Low Lung Function in Young Adult Life Is Associated with Early Mortality. *Am J Respir Crit Care Med* 2017; 195: 1399-1401.

58. Agusti A, Noell G, Brugada J, Faner R. Lung function in early adulthood and health in later life: a transgenerational cohort analysis. *Lancet Respir Med* 2017; 5: 935-945.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

LEGENDS FOR FIGURES

Figure 1. Trajectories of 5 wheeze classes obtained with Partition-Around-Medoids (PAM) algorithm: percentage of participants with reported wheezing in each time interval in the 5 cohorts

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)

Figure 2. Comparison of internal homogeneity of wheezing phenotypes derived using the spells Partition-Around-Medoids (PAM) (panels a and b) and binary LCA approaches (panels c and d) among 7719 subjects with complete data on wheezing from infancy to adolescence

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)

*For Intermittent, 0=No wheeze, 1=single spell, 2=intermittent spells

Plots b) and d) are multi-dimensional heatmaps, which show the density of the distribution of each of the six derived variables, each of which are represented as a row. The scale of the variables (quantitative and categorical) is shown at the bottom on the plot. The segments in the top bar represent each cluster and their relative sizes. The distribution of each indicator within each cluster is shown vertically. In Figure 2b, for example, intermittent spells (as represented by category 2 for the Intermittent variable) is only present in the pink INT cluster; in the LCA model (2d), intermittent wheeze is present in all classes.

Figure 3. The proportion of study participants with asthma diagnosis in adolescence in each Partition-Around-Medoids (PAM) wheeze phenotype (panel a) and the proportion of subjects with asthma diagnosis in adolescence belonging to each PAM phenotype (panel b).

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)

Figure 4. Proportion of children with allergic sensitization in each Partition-Around-Medoids (PAM) wheeze cluster (Manchester Asthma and Allergy Study)

Never wheeze (NWZ); Early transient (ETW); (3) Late onset (LOW); (4) Persistent (PEW); Intermittent (Int)

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure 5. Lung function trajectories from early school age to early adulthood in MAAS (a) and ALSPAC (b)

Never wheeze (NWZ); Early transient (ETW); (3) Late onset (LOW); (4) Persistent (PEW); Intermittent (INT)

Figure 6. Forest plots of associations of 17q12-21 SNPs (a-c) and CDHR3 (d) with Partition-Around-Medoids (PAM) wheeze clusters

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table 1. Associations of wheezing phenotypes with asthma-related outcomes in adolescence: results from multinomial logistic regression using children with 2+ observations on wheeze (reference class: No wheeze) using weighted membership probabilities. Weights derived from probabilities of class membership across 10 imputation samples from the PAM model. Results are reported as adjusted odds ratios with 95% confidence intervals.

* Models adjusted for maternal history of asthma (recruitment), maternal smoking (recruitment), and low birth weight; ** *Available at the latest follow−up (18 years in IOW, 16 years in MAAS, 15 years in SEATON, 15 years in ASHFORD and 15 years in ALSPAC);* †Sex-, age-, and height-adjusted standard deviation units; FEV_1 = forced expiratory volume in 1 second; FVC = forced vital capacity.

Associations with asthma in adolescence*								
	Current asthma**		**Asthma ever**		**Current** asthma medication**		**Asthma medication ever**	
No wheeze	Reference		Reference		Reference		Reference	
Early transient	2.44	[1.84,3.24]	4.00	[3.45,4.63]	1.96	[1.48,2.58]	3.31	[2.92,3.75]
P value	**<.0001**		**<.0001**		**<.0001**		**<.0001**	
Intermittent	27.06	[20.44,35.84]	22.77	[18.23,28.44]	17.34	[13.17,22.84]	18.47	[15.21,22.43]
P value	**<.0001**		**<.0001**		**<.0001**		**<.0001**	
Persistent	37.72	[29.13,48.85]	48.34	[38.47,60.74]	26.78	[20.90,34.32]	38.97	[32.15,47.24]
P value	**<.0001**		**<.0001**		**<.0001**		**<.0001**	
Late onset	35.44	[27.30,46.00]	17.8	[14.58,21.73]	16.78	[12.96,21.72]	22.32	[18.26,27.27]
P value	**<.0001**		**<.0001**		**<.0001**		**<.0001**	
Associations with lung function in adolescence*								
	Z Scores for FEV1†		**Z Scores for FVC†**		**Z Scores for FEV1/FVC†**			
No wheeze	Reference		Reference		Reference			
Early transient	-0.103	[-0.19,-0.02]	-0.014	[-0.10,0.07]	-0.151	[-0.24,-0.07]		

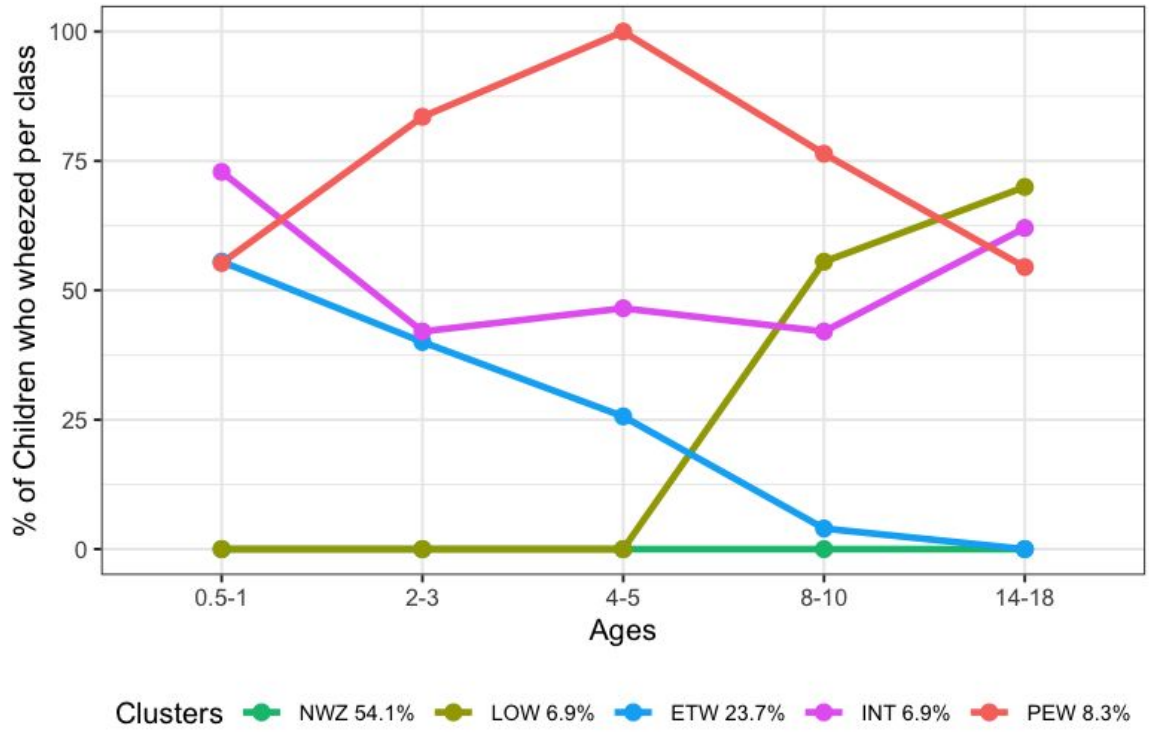
AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

P value	**0.021**		0.748		**<0.0001**
Intermittent	-0.168	[-0.29,-0.05]	0.054	[-0.06,0.17]	-0.379	[-0.49,-0.27]
P value	**0.005**		0.37		**<0.0001**
Persistent	-0.326	[-0.45,-0.20]	0.079	[-0.05,0.21]	-0.707	[-0.83,-0.59]
P value	**<0.0001**		0.221		**<0.0001**
Late onset	-0.003	[-0.13,0.13]	0.159	[0.03,0.29]	-0.302	[-0.43,-0.18]
P value	0.959		**0.015**		**<0.0001**

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure 1. Trajectories of 5 wheeze classes obtained with Partition-Around-Medoids (PAM) algorithm: percentage of participants with reported wheezing in each time interval in the 5 cohorts

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure 2. Comparison of internal homogeneity of wheezing phenotypes derived using the spells Partition-Around-Medoids (PAM) (panels a and b) and binary LCA approaches (panels c and d) among 7719 subjects with complete data on wheezing from infancy to adolescence

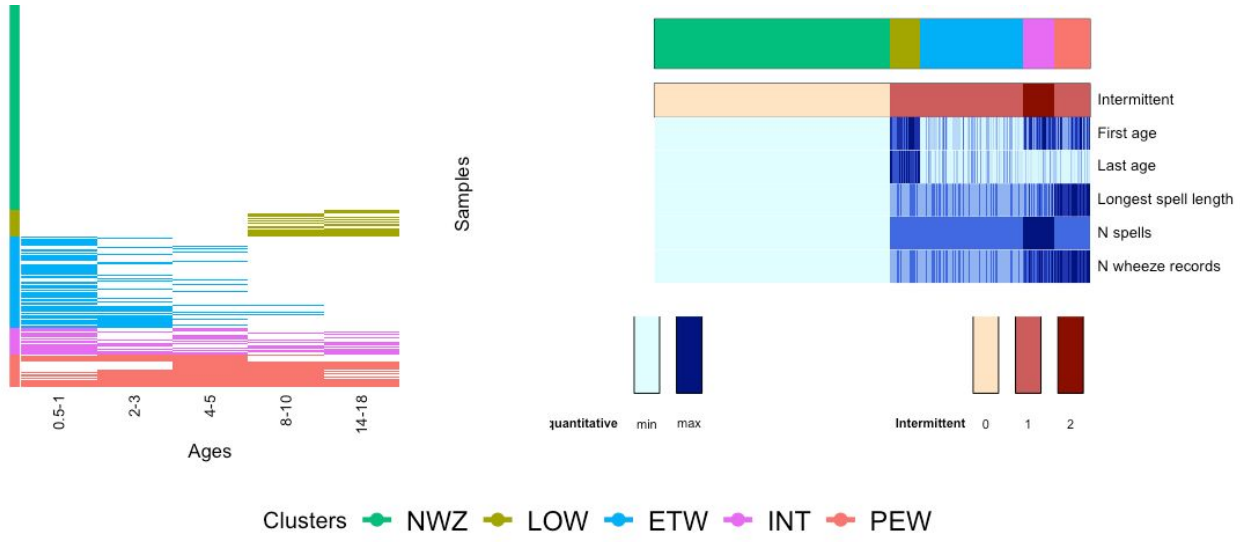
Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)

*For Intermittent, 0=No wheeze, 1=single spell, 2=intermittent spells

Plots b) and d) are multi-dimensional heatmaps, which show the density of the distribution of each of the six derived variables, each of which are represented as a row. The scale of the variables (quantitative and categorical) is shown at the bottom on the plot. The segments in the top bar represent each cluster and their relative sizes. The distribution of each indicator within each cluster is shown vertically. In Figure 2b, for example, intermittent spells (as represented by category 2 for the Intermittent variable) is only present in the pink INT cluster; in the LCA model (2d), intermittent wheeze is present in all classes.

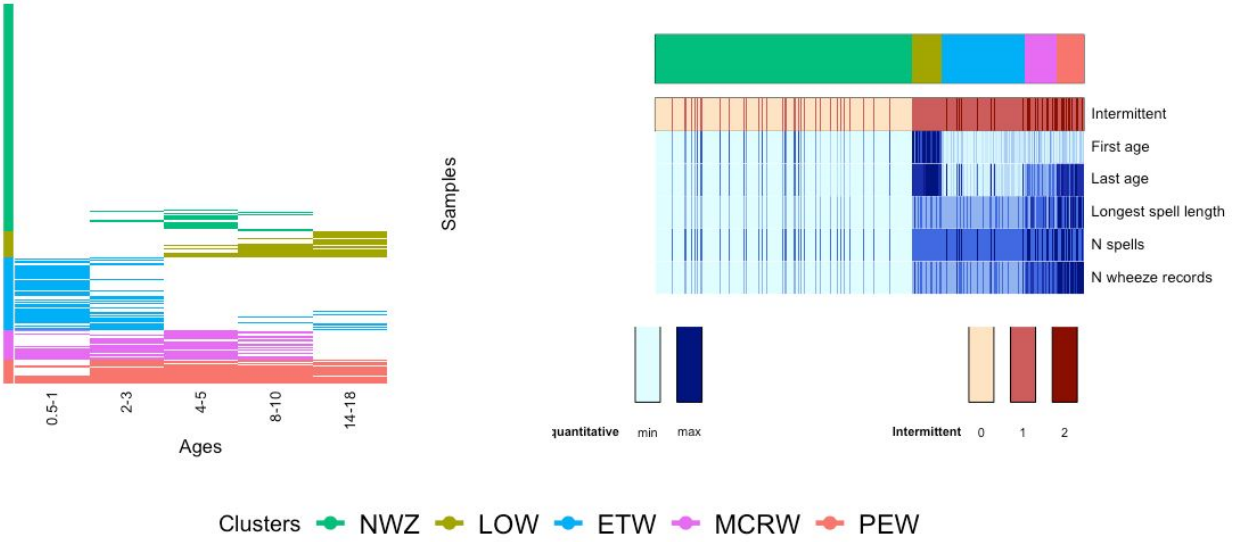
a) PAM: Intra-class individual wheezing patterns

b) PAM: Distribution of multi-dimensional variables by phenotype*



c) LCA: Intra-class individual wheezing patterns

d) LCA: Distribution of multi-dimensional variables by phenotype*

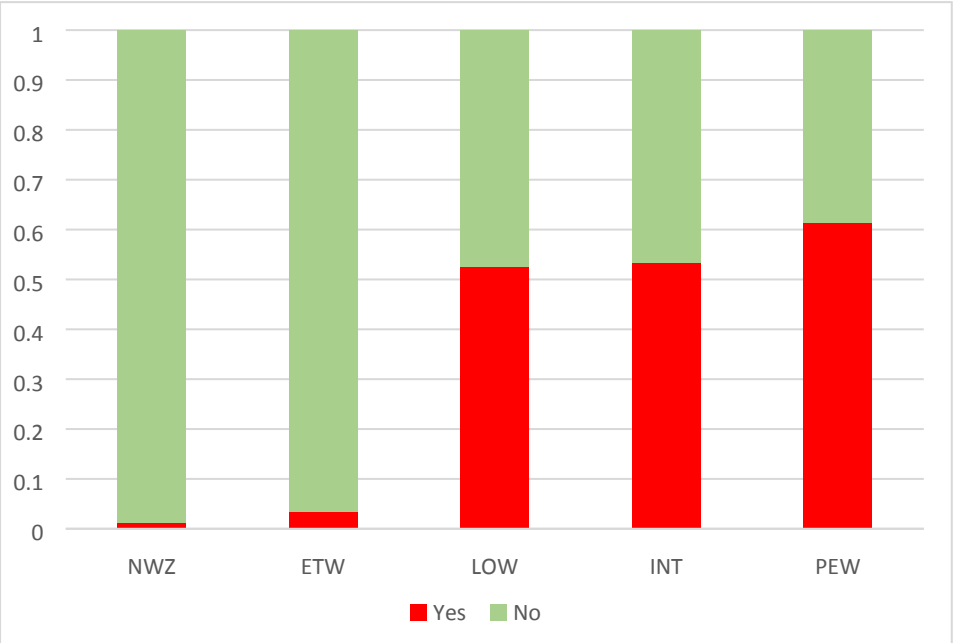


AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

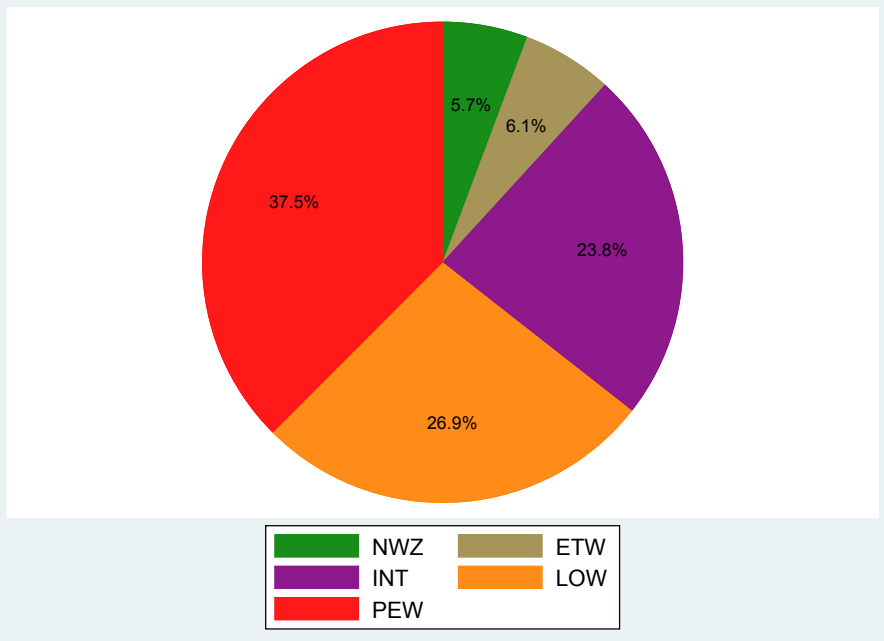
Figure 3. The proportion of study participants with asthma diagnosis in adolescence in each Partition-Around-Medoids (PAM) wheeze phenotype (panel a) and the proportion of subjects with asthma diagnosis in adolescence belonging to each PAM phenotype (panel b).

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)

a)



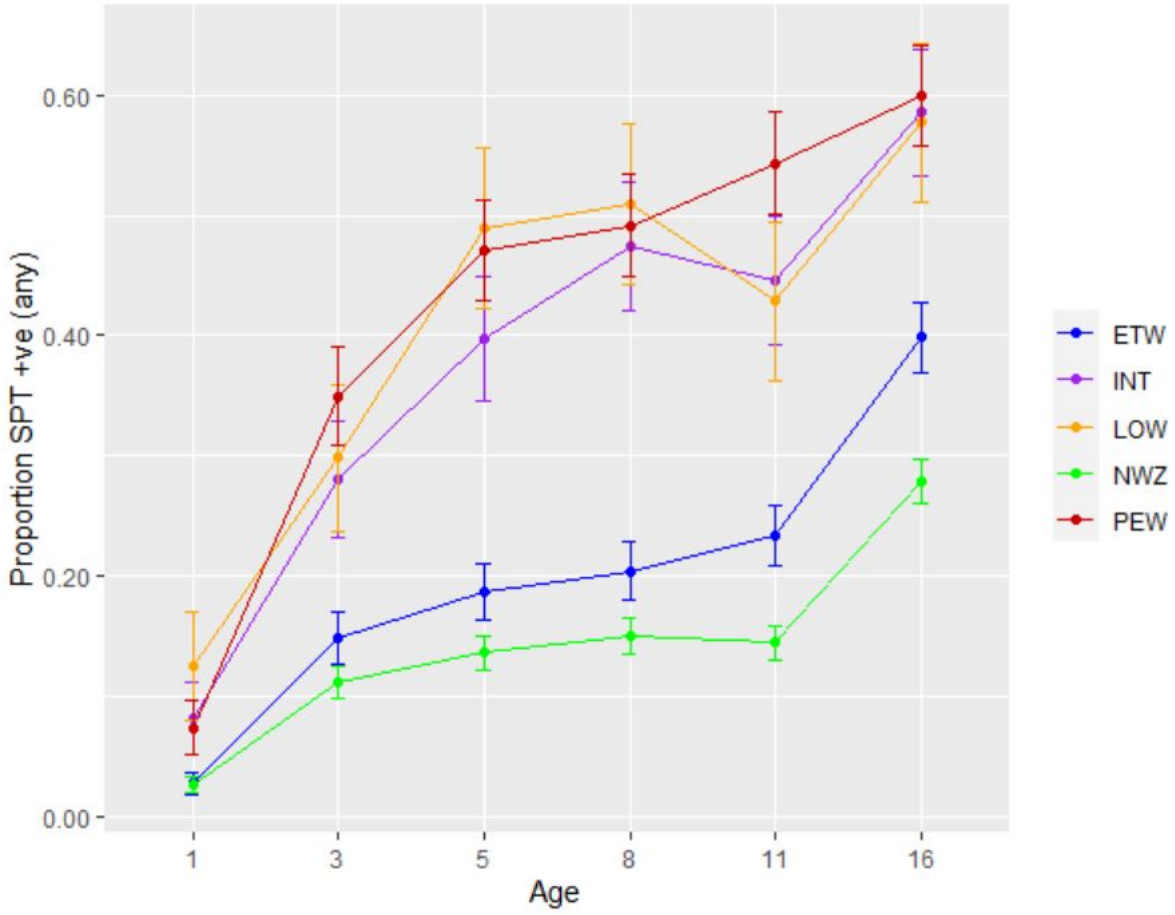
b)



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure 4. Proportion of children with allergic sensitization in each Partition-Around-Medoids (PAM) wheeze cluster (Manchester Asthma and Allergy Study)

Never wheeze (NWZ); Early transient (ETW); (3) Late onset (LOW); (4) Persistent (PEW); Intermittent (Int)

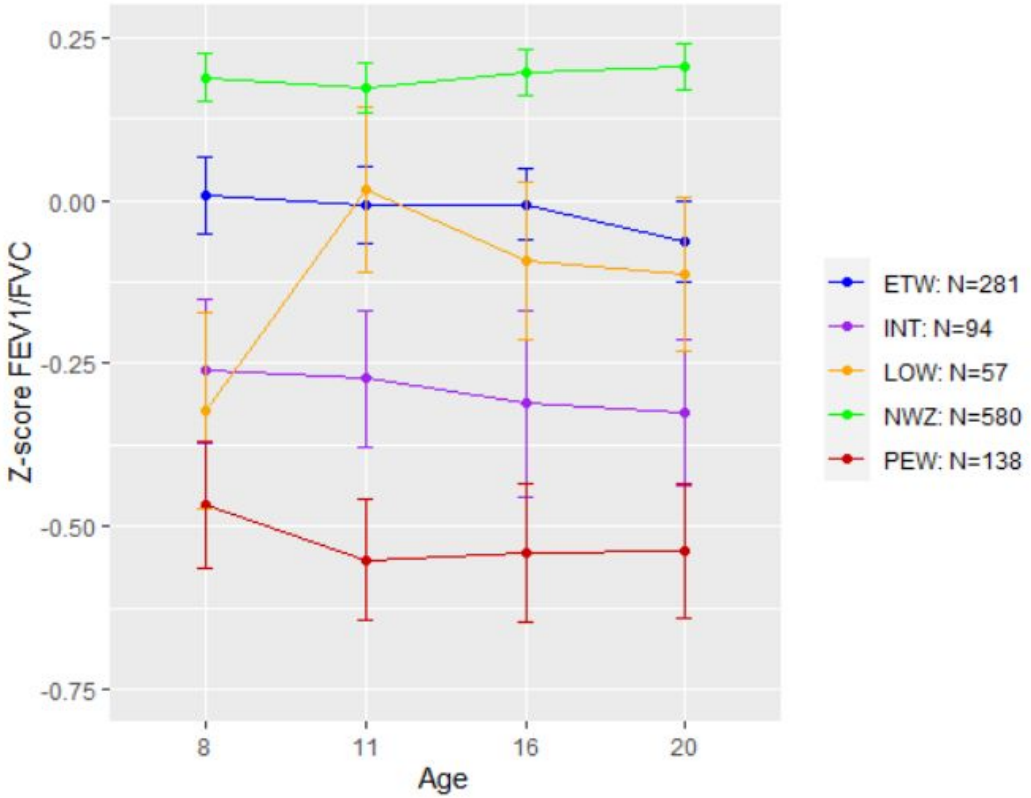


AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

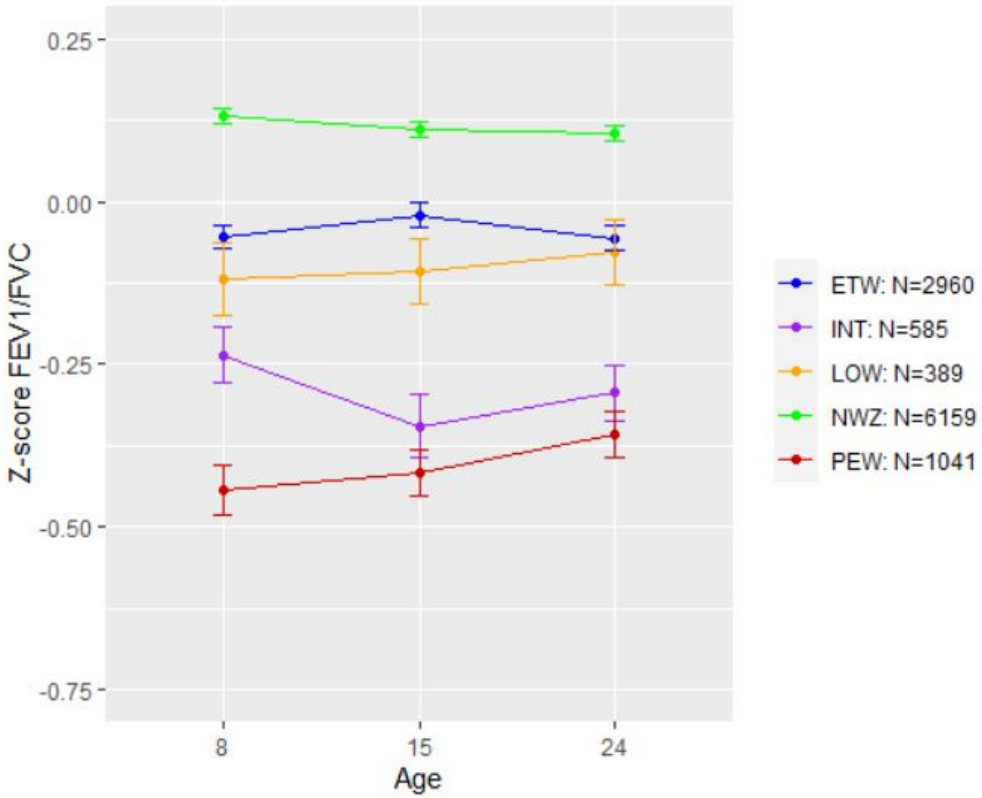
Figure 5. Lung function trajectories from early school age to early adulthood in MAAS (a) and ALSPAC (b)

Never wheeze (NWZ); Early transient (ETW); (3) Late onset (LOW); (4) Persistent (PEW); Intermittent (INT)

a) MAAS



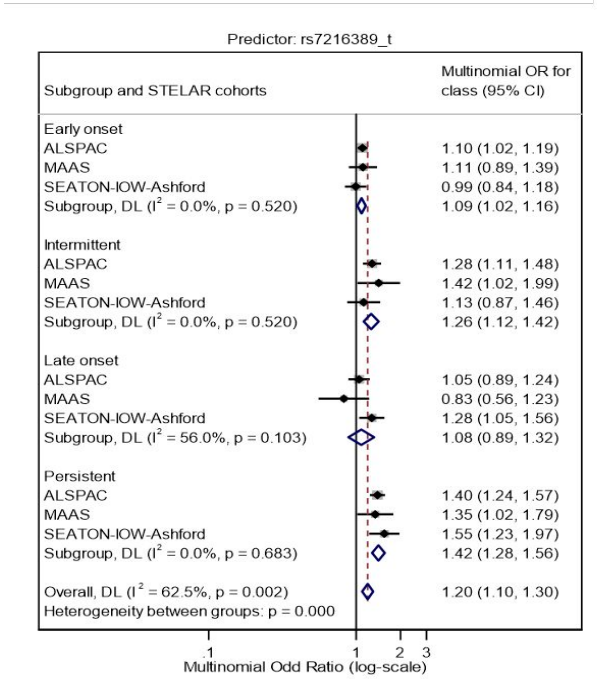
b) ALSPAC



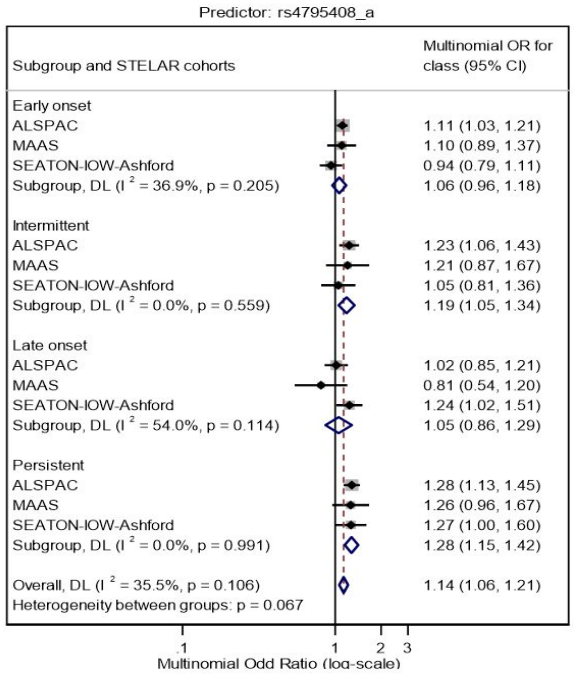
AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure 6. Forest plots of associations of 17q12-21 SNPs (a-c) and CDHR3 (d) with Partition-Around-Medoids (PAM) wheeze clusters

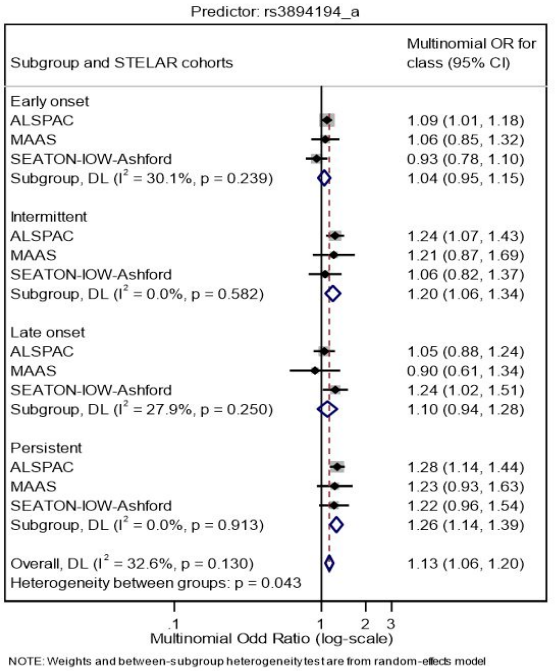
a) rs7216389



b) rs4795408

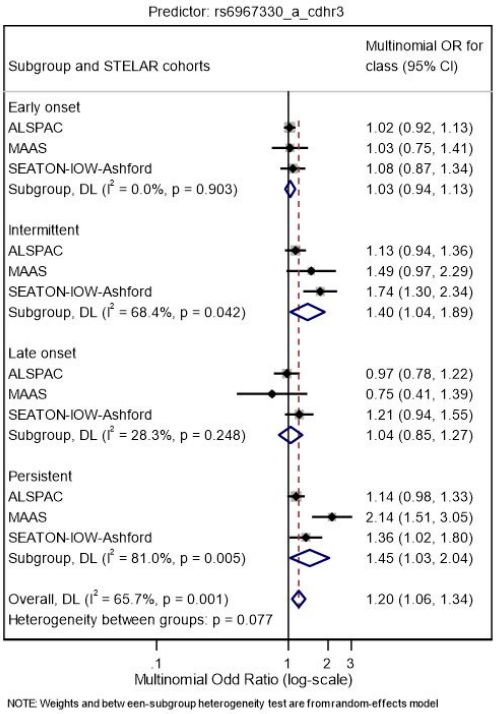


c) rs3894194



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

d) rs697330 (CDHR3)



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

SUPPLEMENTARY APPENDIX

Modelling Wheezing Spells Identifies Phenotypes with Different Outcomes and Genetic Associates

Sadia Haider[1], Raquel Granell[2], John Curtin[3], Sara Fontanella[1], Alex Cucco MSc[1], Stephen Turner[4,5], Angela Simpson[3], Graham Roberts[6,7,8], Clare S Murray[3], John W. Holloway[6,7], Graham Devereux[9], Paul Cullinan[1], Syed Hasan Arshad[7,8,10], Adnan Custovic[1]

[1]National Heart and Lung Institute, Imperial College London, UK

[2]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, UK

[3]Division of Infection, Immunity and Respiratory Medicine, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, UK

[4]Royal Aberdeen Children's Hospital NHS Grampian Aberdeen, AB25 2ZG, UK

[5]Child Health, University of Aberdeen, Aberdeen, UK

[6]Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK

[7] NIHR Southampton Biomedical Research Centre, University Hospitals Southampton NHS Foundation Trust, Southampton, UK

[8]David Hide Asthma and Allergy Research Centre, Isle of Wight, UK

[9]Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK

[10]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

SUPPLEMENTARY INTRODUCTION

Motivation for the approach

We took an inter-disciplinary view by looking at how research in other fields (primarily, the social sciences on poverty dynamics (1-7)) could be applied and develop our knowledge of the longitudinal development of wheeze.

Since the early 1990s, the availability of longitudinal income studies has led to an important shift in the conceptualisation of poverty from a static understanding, in which the cross-sectional prevalence of poverty was compared with non-poverty, to a dynamic one concerned with the duration of spells, the temporal sequencing of poverty, and the extent of persistence and recurrence. Such studies have elucidated that individuals with long or recurrent poverty spells in the past were less likely to escape from poverty, and if they did, they were more vulnerable to experiencing poverty again compared with those with infrequent and short spells. Furthermore, the timing of spells was important, with childhood poverty increasing the risk of detrimental health and social outcomes across the life-course compared with poverty experienced later in life. Studies on mechanisms associated with entries to, exits from and recurrent poverty have informed policies with the aim of safeguarding against re-entry and, therefore, reduce the risk of recurrent poverty (1-7). Inspired by this literature and based on observed patterns of individual trajectories assigned to LCA classes, we developed a set of multi-dimensional variables to describe more holistically the temporal variation of wheeze.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

SUPPLEMENTARY METHODS

DATA SOURCES: DESCRIPTION OF COHORTS

ASHFORD

The Ashford study is an unselected birth cohort study established in 1991 in Ashford, UK (8). It included 642 children born between 1992 and 1993. Participants were recruited prenatally and followed to age 14 years. Detailed standardised questionnaires were administered at each follow-up to collect information on the natural history of asthma and other allergic diseases. Lung function measurements and SPT was carried out at 8 years of age.

The Avon Longitudinal Study of Parents and Children (ALSPAC)

ALSPAC is a birth cohort study established in 1991 in Avon, UK (9, 10). Pregnant women with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled is 14,541. Of these initial pregnancies, there was a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age.

When the oldest children were approximately 7 years of age, an attempt was made to bolster the study with eligible cases who had failed to join originally. As a result, when considering variables collected from the age of seven onwards (and potentially abstracted from obstetric notes) there are data available for more than the 14,541 pregnancies mentioned above. The number of new pregnancies not in the initial sample (known as Phase I enrolment) that are currently represented on the built files and reflecting enrolment status at the age of 24 is 913 (456, 262 and 195 recruited during Phases II, III and IV respectively), resulting in an additional 913 children being enrolled. The total sample size for analyses using any data collected after the age of seven is therefore 15,454 pregnancies, resulting in 15,589 foetuses. Of these 14,901 were alive at 1 year of age.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. The study website contains details of available data through a fully searchable data dictionary and variable search tool: http://www.bristol.ac.uk/alspac/researchers/our-data/

The Manchester Asthma and Allergy Study (MAAS)

MAAS is an unselected birth cohort study established in 1995 in Manchester, UK (11). It consists of a mixed urban-rural population within 50 square miles of South Manchester and Cheshire, located within the maternity catchment area of Wythenshawe and Stepping Hill Hospitals. 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 were born into the study between February 1996 and April 1998. They were followed prospectively for 20 years to date and attended follow-up clinics for assessments, which included lung function measurements, skin prick testing, biological samples (serum, plasma and urine), and questionnaire data collection. The study was approved by the North West – Greater Manchester East Research Ethics Committee.

The Study of Eczema and Asthma to Observe the influence of Nutrition (SEATON)

SEATON is an unselected birth cohort study established in 1997 in Aberdeen, UK, which was designed to explore the relationship between antenatal dietary exposures and asthma outcomes in childhood (12). 2000 healthy pregnant women attending an antenatal clinic, at median 12 weeks gestation, were recruited. An interviewer administered a questionnaire to the women and atopic status was ascertained by skin prick test (SPT). The cohort included 1924 children born between April 1998 and December 1999. Participants were recruited prenatally and followed up by self−completion

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

questionnaire to 15 years of age using postal questionnaires to record the presence of asthma and allergic diseases. The study was approved by the North of Scotland Research Ethics Committee.

The Isle of Wight (IOW) cohort

IOW is an unselected birth cohort study established in 1989 on the Isle of Wight, UK (13-15). After the exclusion of adoptions, perinatal deaths, and refusal for follow-up, written informed consent was obtained from parents to enrol 1,456 newborns (of 1536 born between 1st January 1989 and 28th February 1990). Follow-up-up assessments were conducted to 26 years of age to prospectively study the development of asthma and allergic diseases. At each follow-up, validated questionnaires were completed by the parents. At 10 years, spirometry was performed as described below. Ethics approvals were obtained from the Isle of Wight Local Research Ethics Committee (now named the National Research Ethics Service, NRES Committee South Central – Southampton B) at recruitment and for the subsequent follow-ups.

DEFINITIONS OF VARIABLES (OUTCOMES, DEMOGRAPHIC AND EXPOSURES)

Current wheeze: Defined as a positive response to the question "Has your child had wheezing or whistling in the chest in the last 12 months?" at each follow-up.

Asthma: Current asthma in adolescence was defined as a positive answer to a questions "Has your child had asthma during the past 12 months" at the harmonized point during adolescence.

Information on asthma through childhood was obtained from the responses given to the question "Has your child ever suffered from asthma". Based on the responses, children were divided into two groups: children who had asthma in past (responded "yes" to at least one asthma question) and children who never had asthma (responded "no" to all asthma questions available).

Information about the use of asthma medication during adolescence was obtained from parental reports of whether their child had used any medication and/or received any treatment for asthma in the past 12 months.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Allergic sensitization: Defined as a wheal diameter of 3mm greater than the negative control to one or more allergens.

Parental history of asthma, eczema and hay fever: Defined based on the responses given to the question "have you (and/or your partner) ever had asthma/eczema/hay fever".

Maternal and paternal smoking: Defined based on the response given to the question "do you (or does your partner) smoke", administered during pregnancy.

Low birth weight: Defined as birth weight less than 2500 g based on NHS birth records.

Early−life risk factors were divided into four groups according to timing of exposure; maternal and child characteristics (gender, maternal smoking during pregnancy and maternal history of asthma at recruitment), perinatal (low birth weight adjusted for gestational age), and environmental (pet ownership, smoke exposure after birth).

SPIROMETRY

MAAS: Performed at ages 8, 11, 16 and 20 years according to American Thoracic Society/European Respiratory Society guidelines (16, 17) using a Lilly pneumotachograph system with animated incentive software (Jaeger, Germany). For home visits, we used a flow turbine spirometer (Micro Medical, UK). Subjects were asked to inhale to total lung capacity (TLC), then instructed to perform a forced expiration, through a mouthpiece, to residual volume (RV). The test was repeated at intervals of 30 seconds until 3 technically acceptable traces were obtained. Forced expiratory volume in one second (FEV_1) and Forced vital capacity (FVC) were recorded and the data expressed as FEV_1 % predicted and FEV_1/FVC ratio. Short-acting β2-agonists were withheld for at least four, and long-acting for at least 24 hours prior to testing. Participants were symptom-free at the time of assessment.

ALSPAC: Performed according to American Thoracic Society/European Respiratory Society guidelines (16, 17) using a Vitalograph pneumotachograph system with animated incentive software (Spirotrac, Vitaograph, UK) in a dedicated research clinic by trained technicians. Calibration checks were

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

performed with a standard 3L calibration syringe according to the manufacturer's instructions at the start of each half-day clinic session. Subjects were seated with a nose clip in place and were asked to inhale to TLC, then instructed to perform a forced expiration, through a mouthpiece, to residual volume (RV). The test was repeated at intervals of 30 seconds until 3 technically acceptable traces were obtained from a maximum of eight attempts. FEV_1 and FVC were recorded.

IOW, SEATON and Ashford: Pre-bronchodilator lung function tests were conducted at the follow-ups in adolescence. FVC and FEV1 were measured using a Koko Spirometer and software with a portable desktop device (both PDS Instrumentation, Louisville, KY, USA). Performed according to American Thoracic Society/European Respiratory Society guidelines (16, 17).

Study participants were required to be free of respiratory infection for 2 weeks and not to be taking any oral steroids and were advised to abstain from any β-agonist medication for 6 h.

GENOTYPING AND IMPUTATION

ALSPAC: Participants were genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA) by the Wellcome Trust Sanger Institute (WTSI; Cambridge, UK) and the Laboratory Corporation of America (LCA, Burlington, NC, USA), using support from 23andMe. Haplotypes were estimated using ShapeIT (v2.r644) which uses relationship information to improve phasing accuracy. The phased haplotypes were then imputed to the Haplotype Reference Consortium (HRCr1.1, 2016) panel (18) of approximately 31,000 phased whole genomes. The HRC panel was phased using ShapeIt v2, and the imputation was performed using the Michigan imputation server.

MAAS: Participants were genotyped using the Illumina 610 quad genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA). Prior to imputation samples were excluded on the basis of gender mismatches; minimal or excessive heterozygosity, genotyping call rates of <97%. SNPS were excluded if they had call rates of <95%, minor allele frequencies of <0.5% and HWE p<3x10-8. Prior to

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

imputation each chromosome was pre-phased using EAGLE2 (v2.0.5) (18) as recommended by the Sanger imputation server (19). We then imputed with PBWT (20) with the Haplotype Reference Consortium (release 1.1) of 32,470 reference genomes (19) using the Sanger Imputation Server.

IOW, SEATON and ASHFORD: Participants were genotyped using the Illumina Infinium Omni2.5-8 v1.3 BeadChip genotyping platform (Illumina Inc., San Diego, CA, USA). Genotype QC and imputation was carried out as described for MAAS.

Choice of candidate genes for association analyses

The 17q12-21 and *CDHR3* and SNPs used for this study were chosen based on their previous associations with childhood-onset asthma, either as lead SNPs or associations found in studies which used deep phenotyping.

The first GWAS of asthma reported in 2007 identified multiple markers on chromosome 17q21 as associates of the childhood-onset asthma (21). A comprehensive review which summarised the results of 42 GWASs to date of asthma, different asthma phenotypes and asthma-related traits has been published (22), and provides a summary of the many risk alleles and loci which were replicated in different populations. The most widely replicated asthma locus in GWASs is 17q12-21, hence, we used SNPs from this locus.

A GWAS, which used a specific subtype of early-onset childhood asthma with recurrent, severe exacerbations as an outcome identified a novel gene, Cadherin Related Family Member 3 (CDHR3) as an associate of this specific subtype, but not of doctor-diagnosed asthma (23). Mechanistic studies that followed have suggested that CDHR3 may be a receptor for Rhinovirus C (24).

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

STATISTICAL ANALYSIS

PAM Clustering

PAM is a clustering algorithm that partitions the dataset into a predefined number of clusters and has the advantage of being robust to noise and the presence of outliers. The algorithm selects k-medoid initially and then swaps the medoid object with non-medoid thereby improving the quality of clusters.

The algorithm is based on an iterative procedure that starts with the selection of a representative object for each group. This is called a medoid and represents the most centrally located object within the cluster. Once the medoids have been selected, the remaining objects are assigned to each cluster by minimizing their distance from medoids. The quality of the partition is then measured by the average dissimilarity between an object and the medoid of its cluster. The algorithm selects k-medoids and then swaps each medoid object with a non-medoid thereby improving the quality of clusters.

A key distinction between LCA and PAM is that in our study, the latter does not explicitly model repeated measures, but indicators derived from repeated measures. An advantage of this approach is that excessive variation in the data is "absorbed" whilst retaining important features of change at the individual level. Furthermore, PAM clustering is a simple and flexible algorithm to implement. As it is a non-parametric method, it does not rely on any statistical assumptions and can be used with mixed data types (for example, binary, ordinal, and continuous).

We attempted LCA with our mixed data using STATA's '*gsem*' suite of commands, however, convergence was not achieved. '*POLCA*' package in R does not allow for the simultaneous modelling of categorical and continuous data.

Selection of the optimal number of clusters

With regards to the selection of the optimal number of clusters, the average silhouette width (ASW) has been suggested for finding the number of clusters with PAM (25). It is a simple measurement of cluster quality that does not rely on statistical model assumptions, and is widely used and trusted for

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

comparing the quality of clustering produced by various clustering methods over different numbers of clusters. Furthermore, the silhouette width achieved robust results in the extensive simulation study of Arbelaitz et al. (26). To test the sensitivity of the optimal number of clusters to different indices, we also checked Pearson's Gamma (27), Dunn (27), and Calinski & Harabasz (28) indices. As the results were consistent across all indices, and for brevity, we report the ASW in the manuscript.

Whilst statistical judgements informed the optimal number of classes, we did not rely solely on the ASW, but also visualisations of the internal structure to check for within-class homogeneity, intra-class separation, and guidance from literature on previously derived wheeze clusters. Importantly, clinical judgement was an integral part of the phenotype derivation process.

Sensitivity analysis to determine the stability of the optimal number of phenotypes

We undertook three additional analyses to demonstrate stability of the phenotypes:

1. We ran the Partition-Around-Medoids (PAM) on each single cohort and compared the optimal solution with that of the pooled cohorts.

2. Excluding ALSPAC from the pooled cohort analysis: We excluded ALSPAC cohort from the pooled data and compared the optimal solution for the remaining four cohorts versus inclusion of all five.

3. We investigated cluster stability by running the PAM algorithm on random subsets of data of varying sample sizes, starting with 100% of the data (7719) and reducing it by decrements of 10% until half the sample size was reached. The data were first permuted by ID to ensure that the data was ordered randomly, and for each sample size, the PAM algorithm was run for 10 iterations. We then compared the mean ASW for each sample size over 10 iterations.

The impact of missing data on PAM cluster assignment

We assessed the impact of missing data on cluster assignment in the joint analysis of five cohorts. Wheeze observations were assumed to be missing at random. Based on this assumption, we used the

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

framework of Basagaña *at al.* (29)*,* which integrates multiple imputation (30) (MI) into cluster analysis. MI was applied to the wheeze data using the *ice* suite of commands in Stata 15 (31, 32). Due to the computational intensity of clustering the data, we imputed a maximum of 10 completed data sets, however, 3 to 10 imputed data sets are recommended to obtain reliable results (33). We applied the PAM algorithm to each of the completed data sets and obtained 10 values for the average silhouette index to determine the optimal number of clusters (k_{fin}). k_{fin} was chosen as the mode of the optimal number of clusters across imputed samples. We also examined the distribution of the silhouette index over the data sets by the number of clusters ranging from 2 to 6, and used the median silhouette index over the samples as an additional guide for selecting $k_{fin.}$ We refit the cluster analysis with $k= k_{fin}$ and calculated the membership probability of belonging to each cluster for each child. Children were assigned to the cluster with the largest probability. Finally, we compared individual cluster assignments from the analyses using data from complete cases compared with the imputed datasets using the Adjusted Rand Index.

Comparison of wheezing phenotypes derived using binary LCA and spells PAM approaches

To ascertain the ability of the multi-dimensional indicators to better describe the temporal variation of wheeze developments compared to adopting only presence/absence of symptoms, we compared the clusters obtained with PAM with the phenotypes derived by applying LCA (34) to the binary wheeze data. Detailed description of the joint LCA for the five STELAR cohorts is provided in the manuscript by Oksel et al (35), and is briefly outline below. To facilitate comparison with PAM-derived clusters, in the current analysis participants were assigned to each phenotype according to the maximum posterior probability.

To check for phenotypic homogeneity across cohorts, we stratified class allocation to each cohort by individual wheeze patterns across five time points. We also calculated the ARI for classifications of individual cohort allocations versus the joint cohort allocations.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Latent Class Analysis

To control for cohort−specific variation, Cohort ID was included in the LCA model as an additional predictor by transforming the 5−category variable into a set of four dummy variables and including them as covariates. The largest cohort, ALSPAC, was treated as the non−coded category to which all other cohorts were compared. The expectation maximization algorithm was used to estimate relevant parameters, with 100,000 iterations and 500 replications.

Model Selection: To assess model fit, we used (1) the Bayesian information criterion (BIC), (2) the Akaike information criterion (AIC), (3) Lo−Mendel−Rubin likelihood ratio test (LMR), (4) Bootstrapped likelihood ratio and, (4) quality of classification certainty (model entropy). The BIC is an index used in Bayesian statistics to choose among a set of competing models; the model with the lowest BIC is preferred. Using the lowest BIC as a selection criterion, the best fitting model was chosen as the five-class solution with a nominal covariate (BIC:31340).

LCA vs. PAM

We then compared the within-class homogeneity of both models. We checked the stability of cluster allocations in both models using the ARI, and plotted transitions between classes using alluvial plots. We investigated changes in within-class homogeneity and assessed immutability/mobility of class allocation by cross-classification of phenotypes from both clustering methods. To check for phenotypic homogeneity across cohorts, we stratified class allocation to each cohort by individual wheeze patterns across five time points. We also calculated the ARI for classifications of individual cohort allocations versus the joint cohort allocations. Specifically, we applied PAM and LCA for each cohort separately and compared the final partitions to ascertain whether wheeze patterns assignments were stable in the different cohorts. Furthermore, the adjusted Rand index (ARI) was used to evaluate the agreement between individual assignments using both algorithms, and the stability of assignments when the five cohorts were pooled compared with being modelled singly.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Association of PAM clusters with early−life risk factors and clinical outcomes in adolescence

We used multinomial logistic regression models to ascertain early−life risk factors associated with each PAM-phenotype and examine their relationship with doctor−diagnosed asthma and asthma medication use in adolescence; results are reported as relative risk ratios (RRR), also known as multinomial odds ratios (OR) with 95% confidence intervals (CIs). We tested the validity of derived phenotypes by examining their relationships with asthma and asthma medication use at the last follow-up using logistic regression models. We also investigated associations with lung function outcomes in adolescence (ALSPAC at age 15, IOW at age 18, MAAS at age 16, and SEATON at age 15) using height−, age− and sex−adjusted Z−scores for FEV1, FVC and FEV1/FVC. In MAAS and ALSPAC we derived longitudinal trajectories of lung function for each wheeze cluster. Models were adjusted for potential confounders, including maternal history of asthma, maternal smoking and low birth weight.

Before running multivariable regression analyses, we tested for multicollinearity in a number of ways. Firstly, we ran cross-tabulations for all pairs of categorical independent variables to ensure that there was not a high association between them. We checked collinearity diagnostics, for example, Variance Inflation Factor (VIF) (ensuring that values did not exceed 10), and the stability of estimates and their standard errors by adding variables one at a time.

One-way ANOVA tests were undertaken to ascertain overall statistical differences in FEV_1/FVC means by wheeze phenotypes at each time point in addition to post-hoc pairwise tests corrected for multiple comparisons using Tukey's HSD test.

Association between PAM clusters and genetic variants in *17q12-21* and *CDHR3*: Meta-analysis

We performed clumping (with significance thresholds of 0.05 for index and clumped SNPs) to keep only one representative SNP per linkage disequilibrium block, leaving rs7216389, rs4795408, and rs3894194 in the final analysis. Multinomial logistic regression analysis was used with

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Never/infrequent wheeze as the reference. For SNP's, the additive (dosage) model was used, where the number of risk alleles was treated as a continuous variable in the regression analysis.

The analysis was performed independently in ALSPAC, MAAS and the combined IOW-SEATON-ASHFORD (these 3 cohorts were combined as they were genotyped on the same platform, at the same time, and quality controlled and imputed together), which enabled replication across different studies. For each wheezing phenotype, we obtained the pooled multinomial odds ratio, 95% confidence interval and p-value for the association between each SNP and the phenotype. We used 'metan' command in Stata to derive the pooled effect estimates assuming a random-effect model stratified by wheezing phenotype. We assessed heterogeneity between sub-groups in terms of I-squared statistic.

All statistical analyses were carried out using Stata 14 and R.

Acknowledgements

The Ashford research team are grateful to all the participants and their families for their support over the years and to the many fellow researchers who have contributed to the cohort's follow up.

The ALSPAC research team are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

The MAAS research team thanks study participants and their parents for their continued support and enthusiasm, and greatly appreciate the commitment they have given to the project. We also acknowledge the hard work and dedication of the study teams (post-doctoral scientists, physiologists, research fellows, nurses, technicians, and clerical staff).

The SEATON research team are grateful to all the participants and their families for their support over the years and also to the many fellow researchers who have contributed to the cohort's follow up.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

The IOW research team are grateful to all the participants and their families for their support over the years and also to the many fellow researchers who have contributed to the cohort's follow up.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S1. The time period and size of data included in the analyses

Birth Cohort:	IOW	MAAS	SEATON	ASHFORD	ALSPAC	Total
Year of birth	1989	1995	1997	1992	1991	
Questionnaire	Interviewer-administered	Interviewer-administered	Postal	Interviewer-administered	Postal	
Data collection age (years)	1, 2, 4, 10, 18	1, 3, 5, 8, 16	1, 2, 5, 10, 15	1, 2, 5, 8, 14	½, $2^{1/2}$, $4^{3/4}$, $8^{1/2}$, 14	
N (%) of children with complete data on wheezing at five selected time points	912/1496 (60.1%)	667/1184 (56.3%)	499/1734 (28.8%)	492/642 (76.6%)	5149/12290 (41.9%)	7719/17346 (44.5%)
N (%) of children with >=2 observations on wheezing at five selected time points	1455/1496 (97.3%)	1150/1184 (97.1%)	1489/1734 (85.9%)	620/642 (96.6%)	11134/1290 (87.9%)	15848/17346 (91.4%)
N (%) of children with >=2 observations on wheezing at five selected time points & genetic data	1234/1455 (84.8%)	980/1150 (85.2%)	577/1489 (38.8%)	439/620 (70.8%)	6817/11134 (61.2%)	10047/15848 (63.4%)

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S2. Derivation of indicators

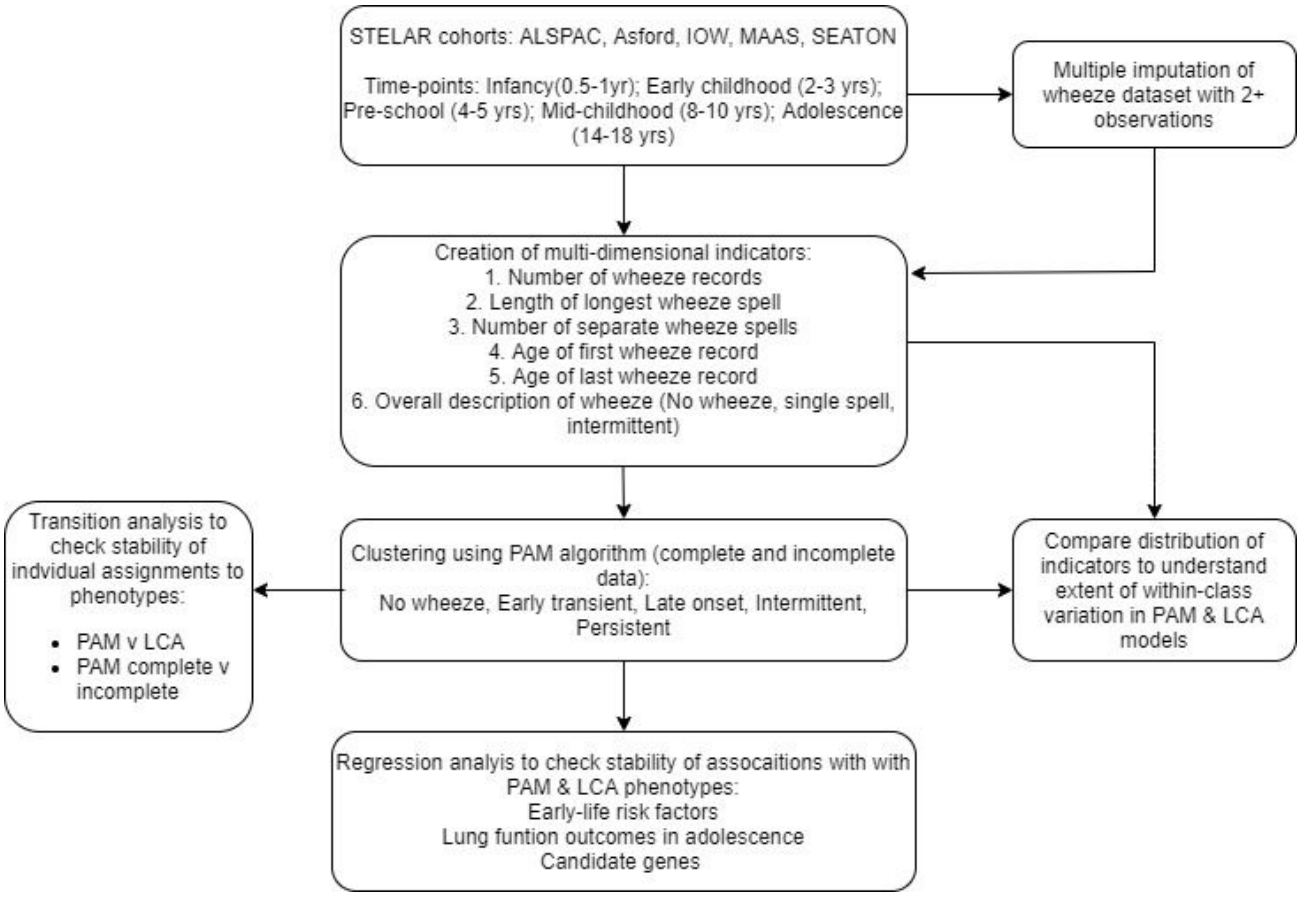
A spell is defined as beginning when wheeze is first observed and ending when non-wheeze is subsequently observed. In the example below, spell lengths can range from one to six consecutive time periods, and individuals can experience multiple spells over the observation period.

For each child, all 6 variables were derived, of which length of the longest spell was one variable. If a child was observed to wheeze at a single time-point (either once only over the observation period or intermittently), the observation for a child with a single spell lasting one time-period was included. We remained agnostic once the variables had been derived, and allowed the PAM algorithm to classify, regardless of duration length or number of spells.

	Wheeze presence/absence						***Derived indicators***					
ID	**TP1**	**TP2**	**TP3**	**TP4**	**TP5**	**TP6**	**Length of longest spell**	**Number of separate spells**	**Number of wheeze observations**	**Spell type**	**Time of wheeze onset**	**Time-point of last wheeze observation**
1	1	1	1	1	1	1	6	1	6	Single	1	6
2	1	0	1	1	1	1	4	2	5	Intermittent	1	6
3	0	1	1	1	1	1	5	1	5	Single	2	6
4	1	0	1	0	1	1	2	3	4	Intermittent	1	6
5	0	0	1	1	1	1	4	1	4	Single	3	6
6	1	0	0	0	1	1	2	2	3	Intermittent	1	6
7	0	0	1	1	0	0	2	1	2	Single	3	4

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure S1. Summary of analysis steps



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S3. The list of SNPs used for this study

*In high linkage disequilibrium with rs4795400

**In high linkage disequilibrium with rs4795406

***From online database SNP Function Prediction (FuncPred): https://snpinfo.niehs.nih.gov/cgi-bin/snpinfo/snpfunc.cgi

SNP	Chromosome	Position	Allele	Risk allele	Nearby Gene***
rs3859192	17	35382174	C/T	T	*GSDM1*
rs3894194	17	35375519	G/A	A	*GSDM1*
rs7216389	17*	35323475	C/T	T	*GSDML*
rs11557467	17	35282160	G/T	G	*ZPBP2*
rs9303277	17	35229995	C/T	C	*IKZF3*
rs2290400	17	35319766	T/C	T	*GSDML*
rs4795405	17	35341943	C/T	T	*ORMDL3*
rs4795408	17**	35361153	A/G	A	*GSDM1*
rs8079416	17	35346239	C/T	C	*ORMDL3*
rs6967330	7	105445687	A/G	A	*CDHR3*

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

SUPPLEMENTARY RESULTS

Table S4. Demographic characteristics of the study population

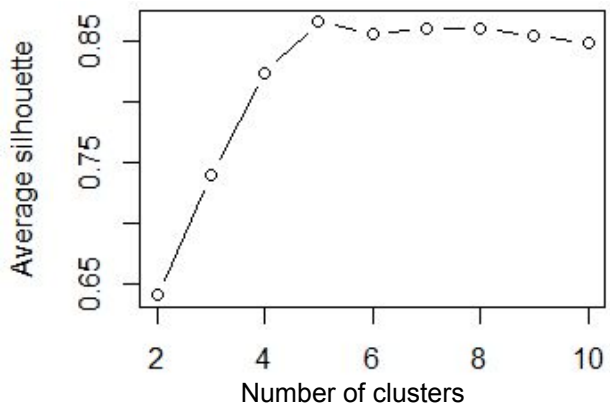
	Children with complete data on wheezing	Children with data on wheezing at 2-4 points
Parental characteristics		
Maternal age at delivery (mean/SD)	29.1 (4.6)	27.6 (5.1)
Maternal asthma ever (recruitment)	12.0% (923/7666)	12.9% (973/7544)
Paternal asthma ever (recruitment)	13.0% (845/6497)	12.4% (700/5639)
Perinatal characteristics		
Male gender	50.4% (3888/7719)	53.0% (4300/8115)
Low birth weight (≤2500 gr)	4.1% (314/7604)	5.6% (433/7750)
Environmental characteristics		
Breastfeeding ever	79.2% (6007/7586)	69.1% (4789/6942)
Maternal smoking (recruitment)	19.2% (1294/6768)	28.7% (1877/6537)
Paternal smoking (recruitment)	27.9% (2123/7618)	36.0% (2742/7642)
Presence of cat (recruitment)	30.4% (2301/7563)	28.6% (2143/7504)
Presence of dog (recruitment)	21.4% (1476/6895)	24.9% (1730/6940)
Outcomes in adolescence (age 14-18)		
Asthma medication ever	29.6% (2116/7153)	34.5% (1579/4580)
Asthma ever	26.5% (1737/6562)	29.2% (828/2837)
Current asthma medication	11.4% (674/5890)	11.4% (256/2245)
Current asthma	12.9% (754/5860)	14.1% (306/2170)
Eczema ever	42.7% (2805/6565)	39.6% (1235/3118)
Cohort		
ALSPAC	46.3% (5149/11134)	53.7% (5985/11134)
MAAS	58.0% (667/1150)	42.0% (483/1150)
SEATON	33.5% (499/1489)	66.5% (990/1489)
IOW	62.7% (912/1455)	37.3% (543/1455)
ASHFORD	79.4% (492/620)	20.6% (128/620)

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S5. Prevalence of wheeze in each cohort and the joint analysis

	Infancy	Early childhood	Pre/early-school	Middle childhood	Adolescence
	0.5-1 years (N/%)	2-3 years (N/%)	4-5 years (N/%)	8-10 years (N/%)	14-18 years (N/%)
ALSPAC	1181	1014	895	660	566
	22.9%	19.7%	17.4%	12.8%	11.0%
ASHFORD	195	124	84	68	71
	39.6%	25.2%	17.1%	13.8%	14.4%
IOW	100	138	188	174	237
	11.0%	15.1%	20.6%	19.1%	26.0%
MAAS	164	141	134	118	107
	24.6%	21.1%	20.1%	17.7%	16.0%
SEATON	122	78	61	66	75
	24.5%	15.6%	12.2%	13.2%	15.0%
JOINT	**1762**	**1495**	**1362**	**1086**	**1056**
	22.8%	**19.4%**	**17.6%**	**14.1%**	**13.7%**

Figure S2. Silhouette plot used to determine the optimal number of clusters in Partition-Around-Medoids (PAM) model



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S6. Percentage of participants assigned to Partition-Around-Medoids (PAM) clusters in single and pooled cohorts' analysis, and adjusted Rand index (ARI) to compare similarity of partitions between each single cohort with the pooled cohorts for PAM & LCA modes.

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)

	ALSPAC	ASHFORD	IOW	MAAS	SEATON	JOINT
NWZ	55.4	44.3	52.1	52.8	56.3	54.1
ETW	24	33.9	17.8	25.2	21	23.7
LOW	7.3	3.9	8.9	2.7	8.4	6.9
INT	6.3	8.1	5.8	9.6	8	6.9
PEW	7	9.8	15.5	9.7	6.2	8.4
	ARI: Single v pooled cohorts					
PAM	0.981	0.953	0.911	0.951	0.995	
LCA	0.827	0.845	0.821	0.846	0.612	

Table S7. Comparison of the average silhouette width across randomly selected subsets of data of different sample sizes; complete pooled cohort data

We investigated cluster stability by running the Partition-Around-Medoids (PAM) algorithm on random subsets of data of varying sample sizes, starting with 100% of the data (7719) and reducing it by decrements of 10% until half the sample size was reached. The algorithm was run for 10 iterations for each sample size. The results demonstrate that the clustering solution is stable with respect to changes in sample size, and five phenotypes were consistently obtained as the optimal solution.

	Number of clusters					
Sample size	2	3	4	5	6	7
50%	0.6465	0.6893	0.7614	**0.7909**	0.7874	0.7867
60%	0.6479	0.6893	0.7620	**0.7916**	0.7881	0.7892
70%	0.6479	0.6893	0.7620	**0.7923**	0.7882	0.7889
80%	0.6483	0.6897	0.7617	**0.7920**	0.7880	0.7891
90%	0.6491	0.6899	0.7616	**0.7915**	0.7875	0.7883
100%	0.6463	0.6874	0.7599	**0.7903**	0.7867	0.7882

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

THE IMPACT OF MISSING DATA ON CLUSTER DERIVATION

Multiple imputation (30) was applied to data from participants with at least two observations of wheeze. We derived 10 completed data sets, and applied PAM clustering to each of these. Table S8 shows the silhouette index for models with 2 to 6 clusters across the imputation samples. The optimal result was 5 clusters in seven data sets and 6 in the remaining three. The highest median silhouette index was for 5 clusters, and this optimal solution was very similar to that derived from a complete data set. Cluster allocation, size, and prevalence of wheeze among 7719 children with complete data in the model using 15,848 individuals with at least two observations are shown in Table S9. Children were assigned with a high degree of certainty, with membership probabilities ranging from 0.80-0.93. Cluster membership certainty was lowest in PEW cluster (0.80).

Figure S3 shows the changes in the allocation of 7719 individuals with complete data from the model using only children with data on wheezing at all five time periods to their most probable class in the model using 15,848 individuals with at least two observations. There was very high agreement between individual cluster assignments from data using complete cases (n=7719) and imputed data (n=15,848), ARI=0.944. This is exemplified in Table S9, which shows cluster allocation and size among 7719 children with complete data in the model using 15,848 individuals with at least 2 observations.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S8. Average silhouette index across 10 multiple imputation samples (joint cohort data)

K=number of clusters

Sample	K2	K3	K4	K5	K6	*kfin*
S1	0.634	0.678	0.748	**0.779**	0.769	5
S2	0.636	0.691	0.748	0.713	**0.769**	6
S3	0.637	0.682	0.750	**0.779**	0.769	5
S4	0.637	0.681	0.748	**0.779**	0.769	5
S5	0.639	0.682	0.751	**0.781**	0.771	5
S6	0.637	0.679	0.749	0.712	**0.769**	6
S7	0.636	0.680	0.749	0.713	**0.769**	6
S8	0.635	0.681	0.750	**0.780**	0.770	5
S9	0.638	0.692	0.748	**0.779**	0.770	5
S10	0.636	0.679	0.748	**0.780**	0.770	5

Table S9. Probability of cluster assignment across 10 multiple imputation samples

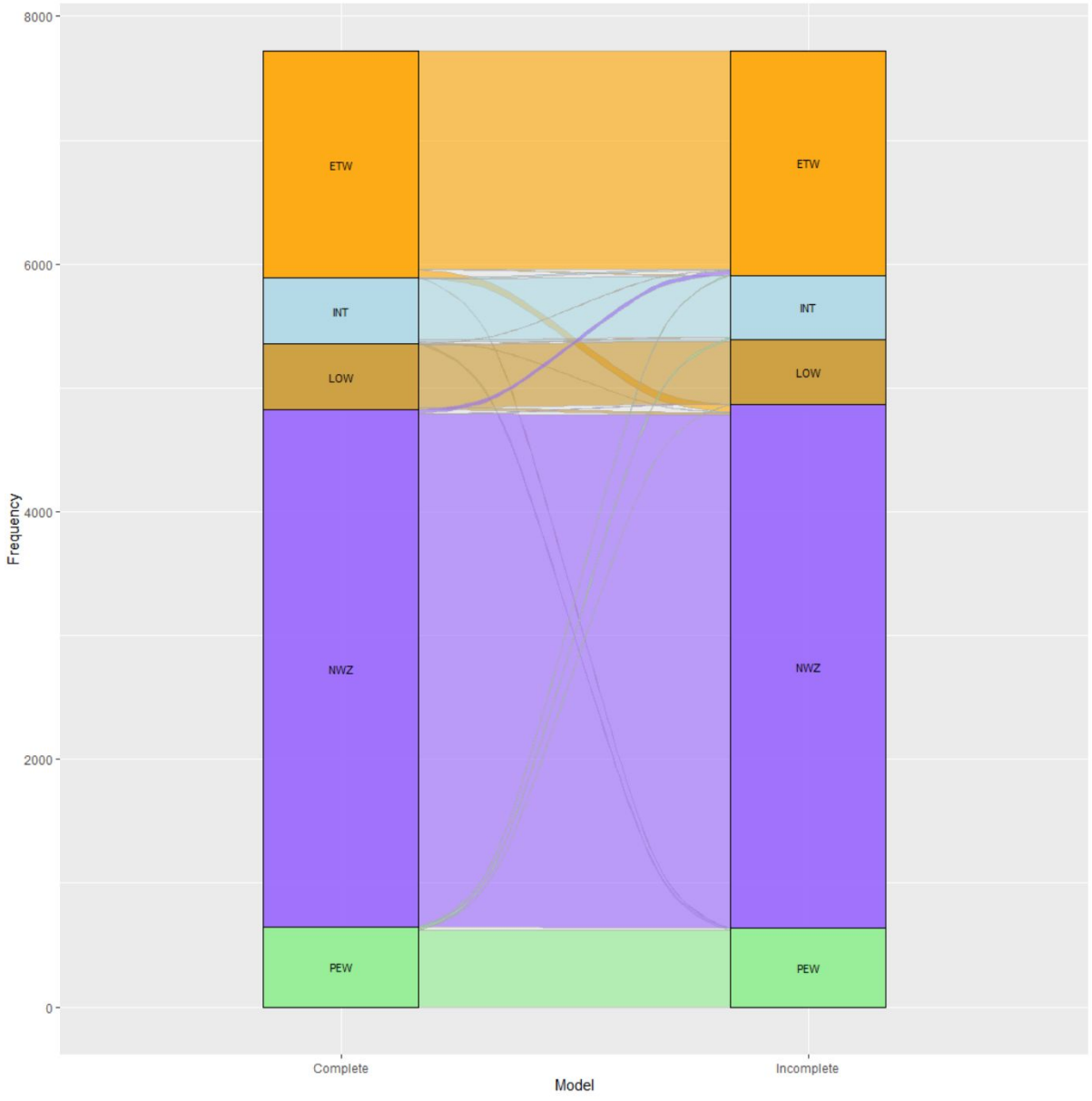
Never wheeze (NWZ); Early transient (ETW); (3) Late onset (LOW); (4) Persistent (PEW); Intermittent (Int)

			Probability of cluster assignment				
Cluster membership		**N/%**	**ETW**	**NWZ**	**INT**	**PEW**	**LOW**
	ETW	3925 (24.8%)	**0.817**	0.003	0.154	0.025	0.000
	NWZ	8821 (55.7%)	0.028	**0.931**	0.006	0.002	0.033
	INT	872 (5.5%)	0.039	0.000	**0.825**	0.127	0.008
	PEW	1463 (9.2%)	0.046	0.000	0.150	**0.795**	0.009
	LOW	767 (4.8%)	0.008	0.002	0.048	0.020	**0.922**
		15,848 (100%)	0.225	0.519	0.103	0.089	0.064

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure S3. Stability of individual allocation to Partition-Around-Medoids (PAM) phenotypes: Alluvial plot shows the transition of phenotype membership for individual participants between models using complete data (n=7719) and the imputed data set from 15,848 individuals with ≥2 observations.

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

COMPARISON OF WHEEZE PHENOTYPES DERIVED USING BINARY LCA AND SPELLS PARTITION-AROUND-MEDOIDS (PAM) APPROACHES

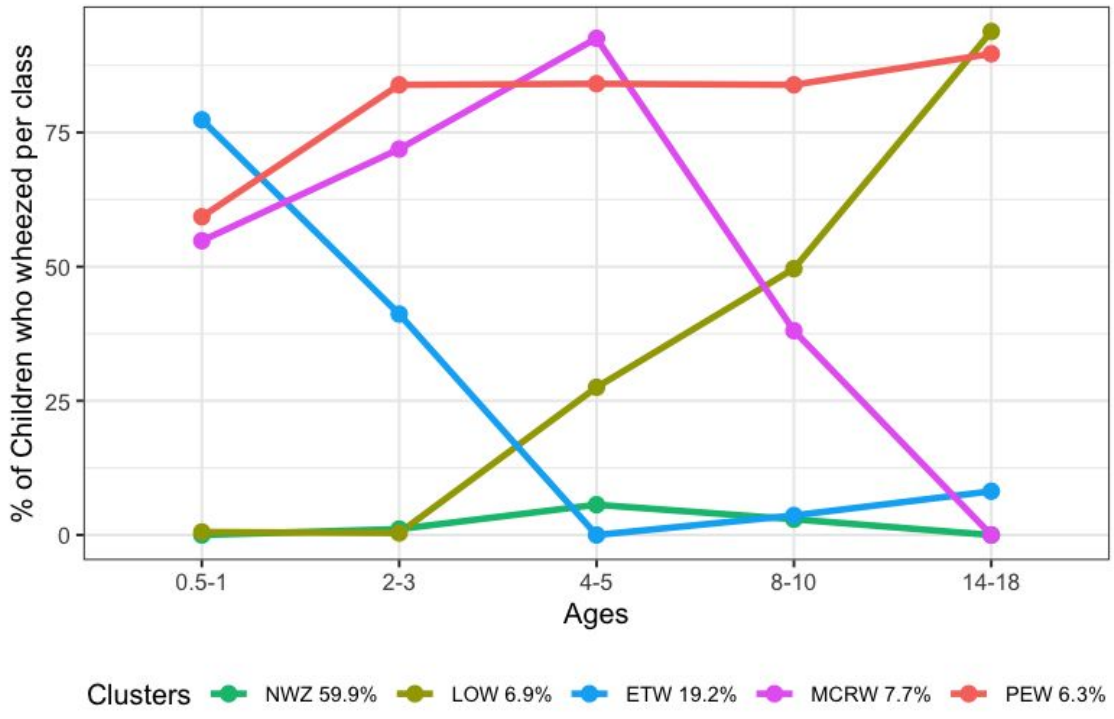
Wheeze phenotypes determined using LCA

Based on statistical fit, a five−class solution was selected as the optimal LCA model(35). To enable compatibility with PAM cluster assignments, we adopted a hard classification, that is we assigned children to their most likely class according to the highest posterior probability of class membership. Based on the onset and duration of wheeze, the classes were labelled as: (1) Never/Infrequent wheeze (59.9%); (2) Early−onset pre−school remitting wheeze (19.2%); (3) Early−onset mid−childhood remitting wheeze (7.7%); (4) Persistent wheeze (6.3%) (5) Late−onset wheeze (6.9%).

Figure S4. Five wheezing phenotypes (latent classes) identified by latent class analysis in 7719 children (infancy: age ½−1, early childhood: age 2−3, pre−school age / early−school age: age 4−5, middle childhood: age 8−10, adolescence: age 14−18). The children were assigned to each phenotype according to the maximum posterior probability.

NWZ: Never/infrequent wheezing; ETW: Early transient (Pre−school remitting) wheezing;

MCRW: Mid−childhood remitting wheezing; PEW: Persistent wheezing; LOW: Late−onset wheezing



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S10. Comparison of descriptive statistics for multi-dimensional indicators by Partition-Around-Medoids (PAM) and LCA classes.

a) Quantitative indicators

		No wheeze		Transient early		Late onset		Persistent		Intermittent	Mid-childhood remitting	Kruskal-Wallis p-value	
		PAM-NWZ	LCA-NWZ	PAM-ETW	LCA-TEW	PAM-LOW	LCA-LOW	PAM-PEW	LCA-PEW	PAM-INT	LCA-MCRW	PAM	LCA
Number of wheeze records	Median	0	0	1	1	1	2	3	4	2	2	<.0001	<.0001
	IQR	[0; 0]	[0; 0]	[1; 2]	[1; 2]	[1; 2]	[1; 2]	[3; 4]	[3; 5]	[2; 3]	[2; 3]		
	Min;Max	[0; 0]	[0; 1]	[1;2]	[1;3]	[1; 2]	[1; 3]	[3; 5]	[2; 5]	[2; 4]	[2; 4]		
Longest spell length	Median	0	1	1	2	1	2	3	4	1	2	<.0001	<.0001
	IQR	[0; 0]	[0; 0]	[1; 2]	[1; 1]	[1; 2]	[1; 2]	[3; 4]	[3; 5]	[1; 2]	[2; 3]		
	Min;Max	[0; 0]	[0; 1]	[1; 2]	[1; 2]	[1; 2]	[1; 3]	[3; 5]	[1; 5]	[1; 3]	[1; 4]		
Number of spells	Median	0	1	1	1	1	1	1	1	2	1	<.0001	<.0001
	IQR	[0; 0]	[0; 0]	[1; 1]	[1; 1]	[1; 1]	[1; 1]	[1; 1]	[1; 2]	[2; 2]	[1; 1]		
	Min;Max	[0; 0]	[0; 1]	[1; 1]	[1; 2]	[1; 1]	[1; 2]	[1; 1]	[1; 3]	[2; 3]	[1; 2]		
First observed age of wheeze	Median		NA	1	0.5	10	10	1	1	1	1	<.0001	<.0001
	IQR	NA	NA	[0.5; 2.5]	[0.5; 1]	[8.5; 14]	[4.75; 14]	[0.5; 2.5]	[0.5; 2.5]	[0.5; 2]	[0.5; 2.5]		
	Min;Max		[NA; 10]	[0.5; 5]	[0.5; 14]	[8; 18]	[1; 18]	[0.5; 5]	[0.5; 5]	[0.5; 5]	[0.5; 5]		
Last observed age of wheeze	Median		NA	2.5	1	14	14	14	14	14	4.75	<.0001	<.0001
	IQR	NA	NA	[0.5; 4]	[0.5; 2.5]	[10; 18]	[14; 18]	[8; 14]	[14; 15]	[8.5; 14]	[4.75; 8.5]		
	Min;Max		[NA; 10]	[0.5; 10]	[0.5; 16]	[8; 18]	[10; 18]	[4; 18]	[8; 18]	[4; 18]	[4; 10]		

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

B) Categorical indicator

		No wheeze		Transient early		Late onset		Persistent		Intermittent	Mid-childhood remitting	Chi-square p-value	
		PAM-NWZ	LCA-NWZ	PAM-ETW	LCA-TEW	PAM-LOW	LCA-LOW	PAM-PEW	LCA-PEW	PAM-INT	LCA-MCRW	PAM	LCA
Intermittent	No episodes												
	N	0	0	0	0	0	0	0	0	0	0	<0.0001	<0.0001
	row % per model	100	100	0	0	0	0	0	0	0	0		
	Single spell												
	N	0	448	1829	1324	533	486	644	303	0	445		
	row % per model	0	15	61	44	18	16	21	10	0	15		
	Intermittent												
	N	0	0	0	160	0	48	0	181	535	146		
	row % per-model	0	0	0	30	0	9	0	34	100	27		

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S11. Comparisons of classification of individual wheeze sequences using PAM and LCA. Green dot=no wheeze; red=wheeze

Wheeze patterns/Age					PAM						LCA					
0.5-1	2-3	4-5	8-10	14-18	ALSPAC	Ashford	IOW	MAAS	SEATON	N classifications	ALSPAC	Ashford	IOW	MAAS	SEATON	N classifications
🟢	🟢	🔴	🟢	🟢	ETW	ETW	ETW	ETW	ETW	1	NWZ	NWZ	NWZ	NWZ	NWZ	1
🟢	🟢	🔴	🔴	🟢	ETW	ETW	ETW	ETW	ETW	1	MCRW	MCRW	LOW	MCRW	LOW	2
🟢	🔴	🟢	🟢	🟢	ETW	ETW	ETW	ETW	ETW	1	ETW	ETW	NWZ	NWZ	ETW	2
🟢	🔴	🔴	🟢	🟢	ETW	ETW	ETW	ETW	ETW	1	MCRW	MCRW	MCRW	MCRW	MCRW	1
🔴	🟢	🟢	🟢	🟢	ETW	ETW	ETW	ETW	ETW	1	ETW	ETW	ETW	ETW	ETW	1
🔴	🔴	🟢	🟢	🟢	ETW	ETW	ETW	ETW	ETW	1	ETW	ETW	ETW	ETW	ETW	1
🟢	🟢	🔴	🟢	🔴	INT	INT	INT	INT	INT	1	LOW	LOW	LOW	LOW	LOW	1
🟢	🔴	🟢	🟢	🔴	INT	INT	INT	INT	INT	1	ETW	ETW	LOW	ETW	ETW	2
🟢	🔴	🟢	🔴	🟢	INT	INT	INT	INT	INT	1	MCRW	ETW	MCRW	PEW	ETW	3
🟢	🔴	🟢	🔴	🔴	INT	NA	INT	INT	INT	1	PEW	NA	PEW	PEW	PEW	1
🟢	🔴	🔴	🟢	🔴	INT	INT	INT	INT	NA	1	PEW	PEW	PEW	PEW	NA	1
🔴	🟢	🟢	🟢	🔴	INT	INT	INT	INT	INT	1	ETW	ETW	LOW	ETW	ETW	1
🔴	🟢	🟢	🔴	🟢	INT	INT	INT	INT	INT	1	ETW	ETW	ETW	ETW	ETW	1
🔴	🟢	🟢	🔴	🔴	INT	INT	INT	INT	INT	1	PEW	PEW	LOW	PEW	PEW	2
🔴	🟢	🔴	🟢	🟢	INT	INT	INT	INT	INT	1	MCRW	MCRW	MCRW	MCRW	MCRW	1
🔴	🟢	🔴	🟢	🔴	INT	INT	INT	INT	INT	1	PEW	PEW	PEW	PEW	PEW	1
🔴	🟢	🔴	🔴	🟢	INT	INT	NA	INT	INT	1	MCRW	MCRW	NA	MCRW	MCRW	1
🔴	🟢	🔴	🔴	🔴	INT	INT	INT	INT	INT	1	PEW	PEW	PEW	PEW	PEW	1
🔴	🔴	🟢	🟢	🔴	INT	INT	NA	INT	INT	1	ETW	ETW	NA	ETW	ETW	1
🔴	🔴	🟢	🔴	🟢	INT	INT	INT	INT	INT	1	MCRW	ETW	MCRW	PEW	ETW	3
🔴	🔴	🟢	🔴	🔴	INT	INT	INT	INT	INT	1	PEW	PEW	PEW	PEW	PEW	1
🔴	🔴	🔴	🟢	🔴	INT	INT	INT	INT	INT	1	PEW	PEW	PEW	PEW	PEW	1
🟢	🟢	🟢	🟢	🔴	LOW	LOW	LOW	LOW	LOW	1	LOW	ETW	LOW	LOW	LOW	2
🟢	🟢	🟢	🔴	🟢	LOW	LOW	LOW	LOW	LOW	1	NWZ	NWZ	LOW	NWZ	NWZ	2
🟢	🟢	🟢	🔴	🔴	LOW	LOW	LOW	LOW	LOW	1	LOW	LOW	LOW	LOW	LOW	1
🟢	🟢	🟢	🟢	🟢	NWZ	NWZ	NWZ	NWZ	NWZ	1	NWZ	NWZ	NWZ	NWZ	NWZ	1
🟢	🟢	🔴	🔴	🔴	PEW	PEW	PEW	PEW	PEW	1	LOW	PEW	LOW	PEW	LOW	2
🟢	🔴	🔴	🔴	🟢	PEW	PEW	PEW	PEW	PEW	1	MCRW	PEW	PEW	PEW	PEW	2
🟢	🔴	🔴	🔴	🔴	PEW	PEW	PEW	PEW	PEW	1	PEW	PEW	PEW	PEW	PEW	1
🔴	🔴	🔴	🟢	🟢	PEW	PEW	PEW	PEW	PEW	1	MCRW	MCRW	MCRW	MCRW	MCRW	1
🔴	🔴	🔴	🔴	🟢	PEW	PEW	PEW	PEW	PEW	1	MCRW	PEW	PEW	PEW	PEW	2
🔴	🔴	🔴	🔴	🔴	PEW	PEW	PEW	PEW	PEW	1	PEW	PEW	PEW	PEW	PEW	1
TOTAL										32						45

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Figure S5. Alluvial plot showing the transitions of wheeze phenotype membership for 7719 individual participants between LCA and Partition-Around-Medoids (PAM)

Never wheeze (NWZ); Early transient (ETW); Late onset (LOW); Persistent (PEW); Intermittent (INT)

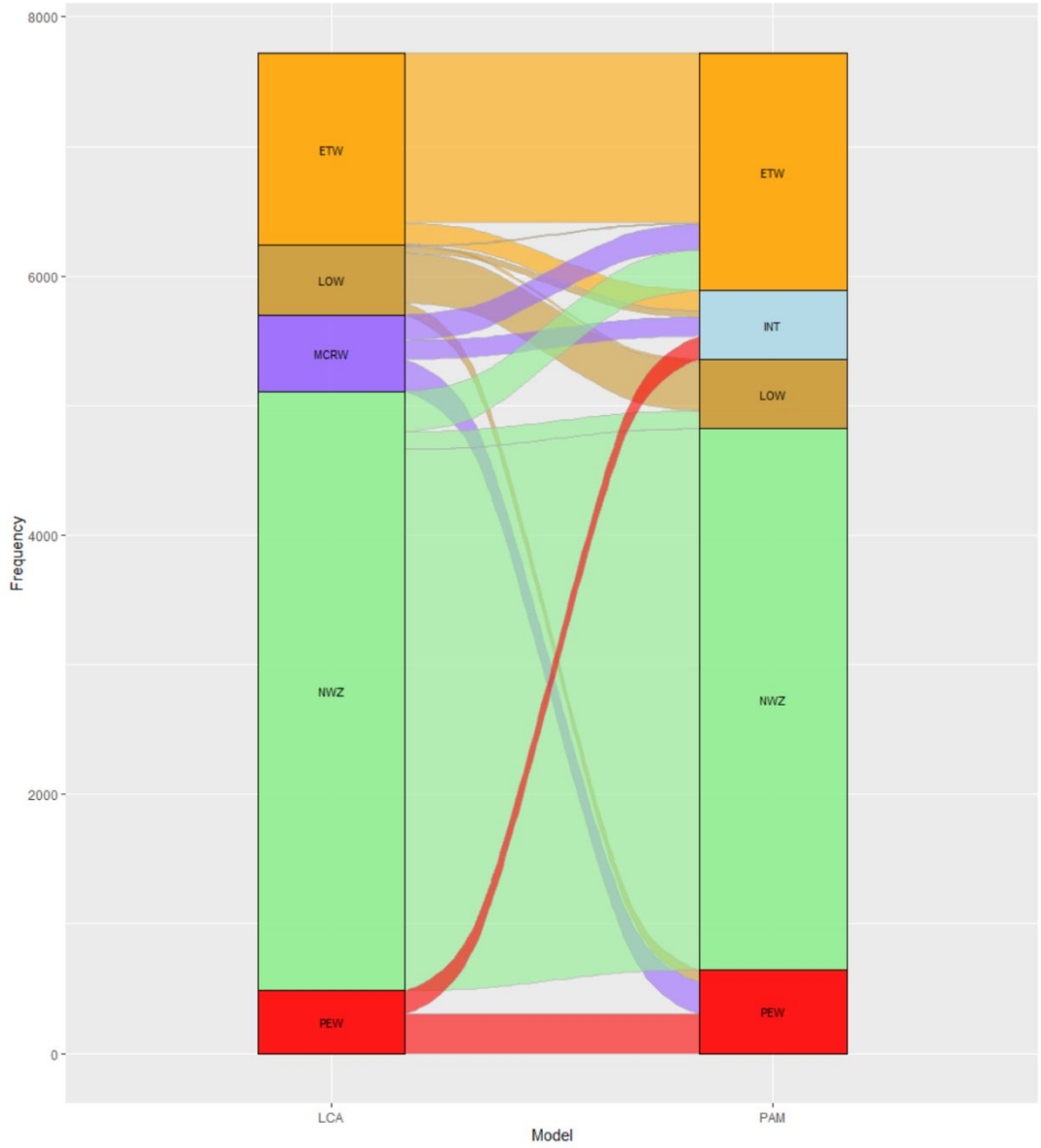
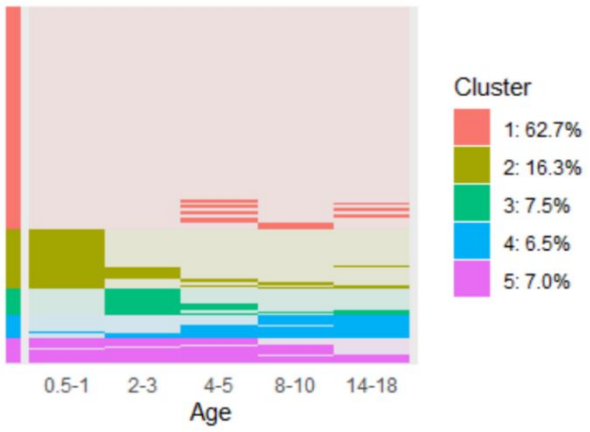


Figure S6. Wheeze phenotypes derived using Partition-Around-Medoids (PAM) algorithm on binary wheeze variables. Plots based on complete sample (N=7719).



AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

ASSOCIATION OF SPELL-BASED PHENOTYPES WITH EARLY–LIFE RISK FACTORS AND ASTHMA-RELATED OUTCOMES DURING CHILDHOOD

Table S12. Associations of wheezing phenotypes with early–life risk factors and skin test responses in mid-school age: results from univariable multinomial logistic regression using children with 2+ wheeze observations (reference class: No wheeze). SPT=skin prick test

		Unadjusted Relative Risk Ratio (95% CI)							
		Early transient		**Intermittent**		**Persistent**		**Late onset**	
Gender (Male)		1.39	[1.29,1.50]	1.49	[1.29,1.71]	1.68	[1.50,1.88]	1.06	[0.92,1.23]
	P value		**<.0001**		**<.0001**		**<.0001**		0.402
Low birth Weight		1.32	[1.10,1.57]	1.45	[1.06,1.97]	1.81	[1.45,2.28]	1.05	[0.73,1.51]
	P value		**0.002**		**0.020**		**<.0001**		0.797
Mother Smoking		1.41	[1.29,1.55]	1.17	[0.97,1.40]	1.44	[1.25,1.65]	1.08	[0.90,1.30]
	P value		**<.0001**		0.096		**<.0001**		0.422
Pet at home		1.11	[1.03,1.20]	0.92	[0.79,1.06]	1.10	[0.98,1.23]	1.00	[0.86,1.16]
	P value		**0.008**		0.234		0.112		0.983
Mother− asthma		1.64	[1.46,1.85]	2.65	[2.20,3.19]	2.96	[2.56,3.42]	1.89	[1.53,2.33]
	P value		**<.0001**		**<.0001**		**<.0001**		**<.0001**
Mother− eczema		1.25	[1.13,1.37]	1.51	[1.27,1.78]	1.82	[1.60,2.07]	1.22	[1.02,1.46]
	P value		**<.0001**		**<.0001**		**<.0001**		**0.032**
Mother− hay fever		1.23	[1.13,1.34]	1.61	[1.38,1.88]	1.99	[1.77,2.23]	1.23	[1.04,1.44]
	P value		**<.0001**		**<.0001**		**<.0001**		**0.014**
Father smoking		1.29	[1.19,1.40]	1.32	[1.13,1.53]	1.32	[1.17,1.49]	1.22	[1.04,1.43]
	P value		**<.0001**		**<.0001**		**<.0001**		**0.013**
Father− asthma		1.29	[1.14,1.48]	1.76	[1.42,2.19]	2.07	[1.75,2.44]	1.54	[1.23,1.93]
	P value		**<.0001**		**<.0001**		**<.0001**		**<.0001**
Father−asthma	male only	1.49	[1.25,1.78]	2.08	[1.58,2.76]	2.42	[1.95,3.00]	1.66	[1.23,2.24]
	P value		**<.0001**		**<.0001**		**<.0001**		**0.001**
Father−asthma	female only	1.08	[0.90,1.30]	1.35	[0.99,1.84]	1.53	[1.21,1.95]	1.35	[0.99,1.84]
	P value		0.419		0.060		**<.0001**		0.056
Father− eczema		1.19	[1.05,1.36]	1.27	[1.00,1.61]	1.29	[1.07,1.56]	1.22	[0.96,1.55]
	P value		**0.008**		**0.048**		**0.007**		0.097
Father− hay fever		1.06	[0.97,1.17]	1.20	[1.01,1.44]	1.29	[1.12,1.48]	1.12	[0.94,1.34]
	P value		0.208		**0.042**		**<.0001**		0.218
Allergic sensitization (SPT)									
Cat		1.34	[1.04,1.72]	3.82	[2.83,5.15]	6.71	[5.33,8.45]	4.02	[2.95,5.46]
	P value		**0.021**		**<.0001**		**<.0001**		**<.0001**
House dust mite		1.32	[1.08,1.60]	3.86	[3.04,4.91]	6.11	[5.05,7.40]	4.15	[3.24,5.31]
	P value		**0.005**		**<.0001**		**<.0001**		**<.0001**
Mixed grasses		0.91	[0.76,1.11]	2.34	[1.83,2.99]	3.26	[2.69,3.96]	2.73	[2.14,3.50]
	P value		0.357		**<.0001**		**<.0001**		**<.0001**

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S13. Associations of wheezing phenotypes with early–life risk factors and allergic sensitisation in early-school age: results from multinomial logistic regression in children with 2+ observations on wheeze (reference class: No wheeze) using weighted membership probabilities. Weights derived from probabilities of class membership across 10 imputation samples from the Partition-Around-Medoids (PAM) model. Results are reported as adjusted relative risk ratios with 95% confidence intervals.

	Early transient		Intermittent		Persistent		Late onset	
Maternal and child characteristics (adjusted by each other)								
Gender (Male)	1.42	[1.37,1.47]	1.48	[1.38,1.59]	1.72	[1.62,1.82]	1.03	[0.96,1.11]
P value		**<.0001**		**<.0001**		**<.0001**		0.355
Maternal smoking	1.4	[1.34,1.47]	1.12	[1.03,1.21]	1.41	[1.32,1.50]	1.06	[0.97,1.15]
P value		**<.0001**		**0.007**		**<.0001**		0.192
Maternal history of asthma	1.64	[1.55,1.73]	2.63	[2.41,2.88]	3.20	[2.98,3.43]	1.97	[1.79,2.17]
P value		**<.0001**		**<.0001**		**<.0001**		**<.0001**
Perinatal characteristics adjusted by maternal and child characteristics								
Low birth weight	1.32	[1.21,1.44]	1.35	[1.16,1.58]	1.77	[1.57,1.98]	1.03	[0.87,1.23]
P value		**<.0001**		**<.0001**		**<.0001**		0.713
Environmental characteristics adjusted by maternal, child, perinatal and environmental characteristics								
Cat ownership	0.97	[0.93,1.01]	0.84	[0.78,0.91]	0.89	[0.84,0.95]	1.05	[0.98,1.14]
P value		0.130		**<.0001**		**<.0001**		0.187
Dog ownership	1.11	[1.06,1.16]	0.98	[0.89,1.07]	1.13	[1.06,1.21]	1.10	[1.01,1.20]
P value		**<.0001**		0.618		**0.001**		**0.034**
Father smoking	1.24	[1.19,1.29]	1.37	[1.27,1.48]	1.31	[1.23,1.39]	1.28	[1.18,1.38]
P value		**<.0001**		**<.0001**		**<.0001**		**<.0001**
Sensitization (age 5 to 7 years) adjusted by maternal, child, perinatal and env. characteristics								
Any allergen	1.11	[1.03,1.19]	2.91	[2.62,3.24]	4.21	[3.86,4.59]	3.52	[3.18,3.89]
P value		**0.009**		**<.0001**		**<.0001**		**<.0001**
Cat	1.34	[1.18,1.53]	3.39	[2.89,3.97]	5.75	[5.08,6.51]	3.72	[3.19,4.34]
P value		**<.0001**		**<.0001**		**<.0001**		**<.0001**
House dust mite	1.32	[1.20,1.46]	3.69	[3.25,4.18]	5.51	[4.97,6.10]	3.74	[3.30,4.23]
P value		**<.0001**		**<.0001**		**<.0001**		**<.0001**
Grass	0.89	[0.80,0.98]	2.18	[1.91,2.48]	2.88	[2.59,3.20]	2.45	[2.16,2.78]
P value		**0.019**		**<.0001**		**<.0001**		**<.0001**

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S14. Comparison of mean FEV1/FVC z-scores by wheeze phenotypes at each time point (MAAS & ALSPAC). Tests for pairwise comparisons conducted with Tukey's HSD test to correct for multiple comparisons. Bold figures indicate statistically significant differences at p<0.05.

PHENOTYPE COMPARISON		MAAS			ALSPAC		
		MEANS		MEAN DIFFERENCE	MEANS		MEAN DIFFERENCE
		AGE 8 (ANOVA p<0.0001)			AGE 8 (ANOVA p<0.0001)		
ETW	INT	0.009	-0.261	0.270	-0.055	-0.236	**0.181**
ETW	LOW	0.009	-0.323	0.331	-0.055	-0.119	0.065
ETW	NWZ	0.009	0.188	0.179	-0.055	0.132	**0.186**
ETW	PEW	0.009	-0.468	**0.476**	-0.055	-0.443	**0.388**
INT	LOW	-0.261	-0.323	0.062	-0.236	-0.119	0.117
INT	NWZ	-0.261	0.188	**0.449**	-0.236	0.132	**-0.368**
INT	PEW	-0.261	-0.468	0.207	-0.236	-0.443	**0.207**
LOW	NWZ	-0.323	0.188	**0.510**	-0.119	0.132	**-0.251**
LOW	PEW	-0.323	-0.468	0.145	-0.119	-0.443	**0.324**
NWZ	PEW	0.188	-0.468	**0.655**	0.132	-0.443	**0.574**
		AGE 11 (ANOVA p<0.0001)					
ETW	INT	-0.008	-0.274	0.265			
ETW	LOW	-0.008	0.017	0.025			
ETW	NWZ	-0.008	0.172	0.181			
ETW	PEW	-0.008	-0.551	**0.543**			
INT	LOW	-0.274	0.017	0.291			
INT	NWZ	-0.274	0.172	**0.446**			
INT	PEW	-0.274	-0.551	0.278			
LOW	NWZ	0.017	0.172	0.155			
LOW	PEW	0.017	-0.551	**0.568**			
NWZ	PEW	0.172	-0.551	**0.723**			
		AGE 16 (ANOVA p<0.0001)			AGE 15 (ANOVA p<0.0001)		
ETW	INT	-0.006	-0.312	0.306	-0.021	-0.345	**0.324**
ETW	LOW	-0.006	-0.092	0.086	-0.021	-0.107	0.086
ETW	NWZ	-0.006	0.196	0.202	-0.021	0.112	0.133
ETW	PEW	-0.006	-0.541	**0.535**	-0.021	-0.417	**0.396**
INT	LOW	-0.312	-0.092	0.220	-0.345	-0.107	**0.239**
INT	NWZ	-0.312	0.196	**0.508**	-0.345	0.112	**-0.457**
INT	PEW	-0.312	-0.541	0.229	-0.345	-0.417	0.072
LOW	NWZ	-0.092	0.196	0.288	-0.107	0.112	**-0.218**
LOW	PEW	-0.092	-0.541	**0.449**	-0.107	-0.417	**0.310**
NWZ	PEW	0.196	-0.541	**0.737**	0.112	-0.417	**0.528**
		AGE 20 (ANOVA p<0.0001)			AGE 24 (ANOVA p<0.0001)		
ETW	INT	-0.064	-0.325	0.262	-0.056	-0.295	**0.239**
ETW	LOW	-0.064	-0.113	0.050	-0.056	-0.078	0.022
ETW	NWZ	-0.064	0.205	0.269	-0.056	0.104	0.160
ETW	PEW	-0.064	-0.537	0.474	-0.056	-0.359	**0.303**
INT	LOW	-0.325	-0.113	0.212	-0.295	-0.078	-0.216
INT	NWZ	-0.325	0.205	**0.530**	-0.295	0.104	**0.398**
INT	PEW	-0.325	-0.537	0.212	-0.295	-0.359	0.064
LOW	NWZ	-0.113	0.205	0.318	-0.078	0.104	0.182
LOW	PEW	-0.113	-0.537	0.424	-0.078	-0.359	**0.280**
NWZ	PEW	0.205	-0.537	**0.743**	0.104	-0.359	**0.462**

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

Table S15. Unadjusted and Benjamani-Hochberg (BH) FDR corrected p-values for associations between 17q12-21 SNPs and CDHR3 with Partition-Around-Medoids (PAM) wheeze clusters; a threshold of 0.05 was used for FDR.

		Unadjusted p-values	BH FDR corrected p-values
rs7216389	Early onset	**0.013**	**0.035**
	Intermittent	**0.000**	**0.000**
	Late onset	0.438	0.539
	Persistent	**0.000**	**0.000**
rs4795408	Early onset	0.255	0.371
	Intermittent	**0.005**	**0.016**
	Late onset	0.609	0.650
	Persistent	**0.000**	**0.000**
rs3894194	Early onset	0.393	0.524
	Intermittent	0.066	0.117
	Late onset	0.235	0.371
	Persistent	**0.000**	**0.000**
rs6967330	Early onset	0.493	0.563
	Intermittent	**0.028**	0.064
	Late onset	0.718	0.718
	Persistent	**0.033**	0.066

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

REFERENCES

1. Mendola D, Busetta A. The Importance of Consecutive Spells of Poverty: A Path-Dependent Index of Longitudinal Poverty. *Rev Income Wealth* 2012; 58: 355-374.

2. Bane MJ, Ellwood DT. Slipping into and out of Poverty - the Dynamics of Spells. *J Hum Resour* 1986; 21: 2-23.

3. Fouarge D, Layte R. Welfare regimes and poverty dynamics: The duration and recurrence of poverty spells in Europe. *J Soc Policy* 2005; 34: 407-426.

4. Muffels R, Fouarge, D. and Dekker, R. . Longitudinal poverty and income inequality: a comparative panel study for the Netherlands, Germany and the UK. In: University of Essex C, editor. European Panel Analysis Group (EPAG) Working Paper 1; 1999.

5. Stevens AH. Climbing out of poverty, falling back in - Measuring the persistence of poverty over multiple spells. *J Hum Resour* 1999; 34: 557-588.

6. Mood C. The not-very-rich and the very poor: Poverty persistence and poverty concentration in Sweden. *J Eur Soc Policy* 2015; 25: 316-330.

7. Layte R, Whelan CT. Moving in and out of poverty - The impact of welfare regimes on poverty dynamics in the EU. *Eur Soc* 2003; 5: 167-191.

8. Atkinson W, Harris J, Mills P, Moffat S, White C, Lynch O, Jones M, Cullinan P, Newman Taylor AJ. Domestic aeroallergen exposures among infants in an English town. *Eur Respir J* 1999; 13: 583-589.

9. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Davey Smith G. Cohort Profile: The 'Children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2013; 42: 111-127.

10. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013; 42: 97-110.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

11. Custovic A, Simpson BM, Murray CS, Lowe L, Woodcock A, Asthma NACM, Allergy Study G. The National Asthma Campaign Manchester Asthma and Allergy Study. *Pediatr Allergy Immunol* 2002; 13: 32-37.

12. Martindale S, McNeill G, Devereux G, Campbell D, Russell G, Seaton A. Antioxidant intake in pregnancy in relation to wheeze and eczema in the first two years of life. *Am J Respir Crit Care Med* 2005; 171: 121-128.

13. Kurukulaaratchy RJ, Fenn M, Twiselton R, Matthews S, Arshad SH. The prevalence of asthma and wheezing illnesses amongst 10-year-old schoolchildren. *Respir Med* 2002; 96: 163-169.

14. Kurukulaaratchy RJ, Fenn MH, Waterhouse LM, Matthews SM, Holgate ST, Arshad SH. Characterization of wheezing phenotypes in the first 10 years of life. *Clin Exp Allergy* 2003; 33: 573-578.

15. Arshad SH, Holloway JW, Karmaus W, Zhang H, Ewart S, Mansfield L, Matthews S, Hodgekiss C, Roberts G, Kurukulaaratchy R. Cohort Profile: The Isle Of Wight Whole Population Birth Cohort (IOWBC). *Int J Epidemiol* 2018; 47: 1043-1044i.

16. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Crapo R, Enright P, Van Der Grinten C, Gustafsson P. Standardisation of spirometry. *European respiratory journal* 2005; 26: 319-338.

17. Beydon N, Davis SD, Lombardi E, Allen JL, Arets HG, Aurora P, Bisgaard H, Davis GM, Ducharme FM, Eigen H, Gappa M, Gaultier C, Gustafsson PM, Hall GL, Hantos Z, Healy MJ, Jones MH, Klug B, Lodrup Carlsen KC, McKenzie SA, Marchal F, Mayer OH, Merkus PJ, Morris MG, Oostveen E, Pillow JJ, Seddon PC, Silverman M, Sly PD, Stocks J, Tepper RS, Vilozni D, Wilson NM, American Thoracic Society/European Respiratory Society Working Group on I, Young Children Pulmonary Function T. 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: 1304-1345.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

18. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, A LP. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016; 48: 1443-1448.

19. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki AE, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL, Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, Van den Berg LH, Van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PI, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, Soranzo N, Walter K, Swaroop A, Cucca F, Anderson CA, Myers RM, Boehnke M, McCarthy MI, Durbin R, Haplotype Reference C. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; 48: 1279-1283.

20. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* 2014; 30: 1266-1272.

21. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SA, Wong KC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WO. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007; 448: 470-473.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

22. Kim KW, Ober C. Lessons Learned From GWAS of Asthma. *Allergy Asthma Immun* 2019; 11: 170-187.

23. Bonnelykke K, Sleiman P, Nielsen K, Kreiner-Moller E, Mercader JM, Belgrave D, den Dekker HT, Husby A, Sevelsted A, Faura-Tellez G, Mortensen LJ, Paternoster L, Flaaten R, Molgaard A, Smart DE, Thomsen PF, Rasmussen MA, Bonas-Guarch S, Holst C, Nohr EA, Yadav R, March ME, Blicher T, Lackie PM, Jaddoe VW, Simpson A, Holloway JW, Duijts L, Custovic A, Davies DE, Torrents D, Gupta R, Hollegaard MV, Hougaard DM, Hakonarson H, Bisgaard H. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet* 2014; 46: 51-55.

24. Bochkov YA, Watters K, Ashraf S, Griggs TF, Devries MK, Jackson DJ, Palmenberg AC, Gern JE. 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: 5485-5490.

25. Kaufman LRPJ. Finding groups in data : an introduction to cluster analysis. New York: Wiley; 1990.

26. Arbelaitz O, Gurrutxaga I, Muguerza J, Perez JM, Perona I. An extensive comparative study of cluster validity indices. *Pattern Recogn* 2013; 46: 243-256.

27. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. *J Intell Inf Syst* 2001; 17: 107-145.

28. Caliński T, Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics* 1974; 3: 1-27.

29. Basagana X, Barrera-Gomez J, Benet M, Anto JM, Garcia-Aymerich J. A Framework for Multiple Imputation in Cluster Analysis. *American Journal of Epidemiology* 2013; 177: 718-725.

30. Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons Inc; 2Rev Ed edition (24 Sept. 2002); 2002.

31. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC; 2017.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC

32. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *J Stat Softw* 2011; 45: 1-20.

33. Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.

34. Linzer DA, Lewis JB. poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software* 2011; 42.

35. Oksel C, Granell R, Haider S, Fontanella S, Simpson A, Turner S, Devereux G, Arshad SH, Murray CS, Roberts G, Holloway JW, Cullinan P, Henderson J, Custovic A, Stelar investigators bTi. Distinguishing Wheezing Phenotypes from Infancy to Adolescence. A Pooled Analysis of Five Birth Cohorts. *Ann Am Thorac Soc* 2019; 16: 868-876.

AJRCCM Articles in Press. Published January 20, 2022 as 10.1164/rccm.202108-1821OC