UCLA
UCLA Previously Published Works

Title

Asthma clustering methods: a literature-informed application to the children's health study data

Permalink

https://escholarship.org/uc/item/3sb9t7dn

Journal

Journal of Asthma, 59(7)

ISSN

0277-0903

Authors

Ross, Mindy K
Eckel, Sandrah P
Bui, Alex AT
et al.

Publication Date

2022-07-03

DOI

10.1080/02770903.2021.1923738

Peer reviewed

Asthma Clustering Methods: A Literature-Informed Application to the Children's Health Study Data

Mindy K. Ross, MD[a,*], **Sandrah P. Eckel, PhD**[b,*], **Alex A. T. Bui, PhD**[c], **Frank D. Gilliland, MD, PhD**[b]

[a]Pediatrics, Pediatric Pulmonology, University of California, Los Angeles, Los Angeles, CA, USA

[b]Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[c]Radiology, University of California, Los Angeles, Los Angeles, CA, USA

Abstract

Objective: The heterogeneity of asthma has inspired widespread application of statistical clustering algorithms to a variety of datasets for identification of potentially clinically meaningful phenotypes. There has not been a standardized data analysis approach for asthma clustering, which can affect reproducibility and clinical translation of results. Our objective was to identify common and effective data analysis practices in the asthma clustering literature and apply them to data from a Southern California population-based cohort of schoolchildren with asthma.

Methods: As of January 1, 2020, we reviewed key statistical elements of 77 asthma clustering studies. Guided by the literature, we used 12 input variables and three clustering methods (hierarchical clustering, *k*-medoids, and latent class analysis) to identify clusters in 598 schoolchildren with asthma from the Southern California Children's Health Study (CHS).

Results: Clusters of children identified by latent class analysis were characterized by exhaled nitric oxide, FEV_1/FVC, FEV_1 percent predicted, asthma control and allergy score; and were predictive of control at two year follow up. Clusters from the other two methods were less clinically remarkable, primarily differentiated by sex and race/ethnicity and less predictive of asthma control over time.

Conclusion: Upon review of the asthma phenotyping literature, common approaches of data clustering emerged. When applying these elements to the Children's Health Study data, latent class analysis clusters—represented by exhaled nitric oxide and spirometry measures-had clinical relevance over time.

Address for correspondence: Mindy K. Ross (pediatricpulmonary@mednet.ucla.edu), Pediatrics, Pediatric Pulmonology, University of California, Los Angeles, 10833 Le Conte Ave, Los Angeles, CA 90095, USA.
*Co-first authors.

Introduction

Asthma is one of the most common chronic medical conditions, affecting over six million children in the United States. Uncontrolled pediatric asthma contributes to significant morbidity and economic burden and is a challenge for healthcare practitioners to manage.[1–4] No universal asthma treatment is available, in part due to heterogeneity among patients (e.g., clinical presentation, age of onset, underlying inflammation, social determinants, environmental exposures).[5–7]

This heterogeneity has inspired widespread application of clustering algorithms to identify patterns among adults and children with asthma.[8] The purpose is to identify phenotypes that reflect underlying inflammation, which can help drive treatment choice in practice. Studies have primarily been performed on pre-existing datasets specifically for asthma studies in participants with more severe asthma.[9] Phenotypes have been characterized by different symptoms presentation, age at onset, symptoms, severity, co-morbidities, and body mass index.[10–13] While there are some consistencies across studies, results vary in part because there has not been a standardized methodological approach. It can be a challenge for clinical practitioners to trust, critically evaluate, and translate these studies to their patients.

Data analysis choices impact clustering results/clinical implications and not all clustering methods are appropriate for all datasets, with input variable type (e.g., binary, continuous, categorical) being an important determinant of appropriate methods. Prosperi *et al*[14] demonstrated that differences in data preprocessing (i.e., variable transformation and dimension reduction through variable selection) impacts results. Deliu *et al*[15] provided an overview of relevant methods for asthma clustering studies; but to our knowledge, there has not been detailed review of the statistical methods used in previous data driven asthma clustering studies, to inform the approach from a clinical perspective.

While most cluster analyses have been performed on formal asthma studies, eventually we envision clustering approaches can be applied to electronic health record (i.e., real-world) data. As an interim step, our objective was to perform clustering in existing data from a population-based cohort study, the Southern California Children's Health Study (CHS). The CHS was originally designed to study the effects of air pollution on pediatric respiratory health and lung development, not specifically targeting asthma patients[16] in order to evaluate the transferability of common clustering approaches. We hypothesized that a literature-informed, rigorous data clustering approach would discover clinically-relevant asthma clusters in the CHS.

Methods

Literature review - asthma clustering studies

As of January 1, 2020 we searched PubMed for asthma phenotyping studies using the following search terms: ("Bronchial hyperreactivity" OR "Asthma" OR asthma* OR bronchial hyper* OR respiratory hyper* OR wheez*) AND ("Phenotype" OR phenotyp* OR endotyp*) AND ("Cluster analysis" OR cluster* OR subtyp* OR subgroup* OR sub-typ* OR sub-group*).

Resulting articles were screened for inclusion first based on title and abstract and then full-text review (independently by MKR and SPE, with differences reconciled by discussion).[17] We included original research articles that applied clustering to clinical asthma. We excluded review articles, non-human subject articles, non-English text articles, articles that applied previously developed clusters, articles with non-asthma study populations, articles with non-clinical variables as the only inputs for clustering, and retracted articles. For each article, we recorded information about: participant ages, sample size, variable selection, dataset preparation/processing, variable standardization, missing data approaches, number and type of input variables included (binary, continuous, categorical) as well as primary clustering method, method to determine number of clusters, the final number of clusters, and methods to characterize the selected clusters.

We retrieved 1,091 articles, the earliest from 1980. The 77 articles meeting inclusion criteria (Figure 1)[17] are summarized in Table E1. Of these, 39 (51%) were adult studies, 27 (35%) were pediatric studies, and the remaining 11 (14%) included both children and adults. More than half (47 or 61%) had <500 participants and the largest had 65,254 participants.

The two most cited articles, Haldar *et al* (2008)[8] and Moore *et al* (2010)[9] appeared to influence the data analysis practices of subsequent studies. Haldar *et al* used a two-step approach using a hierarchical clustering dendrogram and then *k*-means clustering. Moore *et al* first identified clusters using an unsupervised method and then used a supervised method to identify key predictors of cluster membership, characterize the clusters, and enable translation of the clusters to a new population. In our review, the primary clustering method was typically unsupervised: hierarchical clustering (n=26), *k*-means/*k*-medoids (n=29), or latent class analysis (n=10). Of articles using basic hierarchical clustering, only 30% completely specified their approach (agglomerative or divisive, dissimilarity matrix, and linkage criterion). Of these approaches, the most commonly specified linkage criterion was Ward's minimum variance (n=19) and the most common dissimilarity matrix was Gower's distance (n=5). Three studies performed supervised clustering, and all used decision trees (rpart in R)[18] to identify subgroups with different risks of asthma exacerbation.

In regard to input variables, on average in the 77 studies, 16 variables were used to perform clustering. Most were selected based on a combination of clinical relevance (n=37) and/or previous studies (n=10) as well as data-informed dimension reduction (e.g., factor analysis, principal components analysis; n=17), model selection (n=3), and/or considering correlations (n=12). For example, Haldar *et al* (2008)[8] convened experts to identify potential input variables as those "measured in clinical practice" and which "contributed to the clinical evaluation of asthma" but were not a "product of the disease process" and also aimed to avoid choosing multiple variables "representative of the same aspect of disease," aided by principal components analysis to reduce the number of variables.

The final number of clusters was 4 on average (range: 2–8). To identify the number of clusters, 36% of studies (n=28) used subjective visual inspection of the dendrogram, sometimes combined with additional criteria (e.g., 30 participants per cluster[19]). All 10 LCA studies used Bayesian information criterion (BIC), Akaike information criterion (AIC), and/or likelihood ratio tests to determine cluster number. Other methods included

the Silhouette width criteria (n=6), pseudo F-statistic[20] (n=4), and Gap statistic[21] (n=2). The number of clusters identified had a positive, but low correlation with sample size (Spearman's R=0.30).

Missing data were most commonly addressed by including only complete cases (n=49). However, 12 studies applied imputation and 7 used a clustering method that could handle missing data. Input variables included mixed types (continuous and binary/categorical) in 73% (n=55) of the studies. Some form of data normalization was used in 40 studies. Common approaches to data normalization were log-transformation of continuous variables, range (0–1 scale) standardization, or z-score standardization.

Study Population and Dataset

For this study, we analyzed 598 children with asthma enrolled in the Southern California Children's Health Study (CHS). The CHS was a longitudinal cohort of children first recruited from Southern California kindergarten and first-grade classrooms in 2002–2003 to study the effects of air pollution on pediatric respiratory health and lung development.[22] We identified children with asthma in this cohort and performed clustering on Year 6 data (2007–2008) when the children were on average 11.6 ±0.86 years old. We related this cluster membership to asthma symptoms at follow-up two years later (Year 8, 2009).

We included children with baseline questionnaires and who reported having ever been physician-diagnosed with asthma in Year 6, as this was the first visit with pulmonary function testing. We excluded children with diagnoses that could potentially present as respiratory symptoms or mimic asthma (e.g., "heart condition"). Each CHS child participant provided informed assent and a parent/guardian provided informed consent. Baseline and annual follow-up questionnaires were completed by the parent/guardian and, starting in Year 8 of the cohort, by the child. Data were originally collected using a protocol approved by the University of Southern California Institutional Review Board and our analyses were conducted under HS-13-00150.

Questionnaire data and lung function measures

Baseline and annual follow-up questionnaires included questions about: assigned sex, race/ethnicity, if the child was born prematurely (and number of weeks premature), whether a doctor had ever diagnosed the child with asthma, the approximate age when the child was diagnosed with asthma, symptoms over the past 12 months, rescue or controller medication usage over the past 12 months, whether a doctor has ever said that the child's biological mother or father had asthma, and if cigarettes/cigars/pipes are smoked inside the child's home.

Pulmonary function tests (PFTs) were conducted by the children under the guidance of trained technicians (ScreenStar, Morgan Scientific Inc, Haverhill, MA USA).[23] Height and weight were measured concurrently. Percent predicted maximal forced expiratory volume in 1 second (FEV_1) were calculated.[24] Fractional exhaled nitric oxide (FeNO) was collected at a 50 ml/s exhalation flow rate (CLD88-SP with DeNOx, EcoMedics, Duernten, Switzerland/Ann Arbor, MI, USA) as described previously.[25,26]

Variable selection and reduction

We selected CHS variables for inclusion in clustering algorithms based on known clinical asthma characteristics, phenotypes, and previous studies (Table 1). To avoid overweighting a single dimension of a participant's characteristics, we checked for correlation in the final set of input variables. To synthesize overlapping concepts and represent typical asthma-related measurements not captured exactly by CHS questions (e.g., asthma control and allergy labs), we created composite variables, described below and in Table 1. The resultant set of input variables was of mixed types (binary, categorical, continuous).

The composite variable "asthma control" was defined by the frequency of asthma symptoms in response to the questions: In the past 12 months, "How many attacks of wheezing has your child had?", "How often, average, has your child's sleep been disturbed due to wheezing?", and "Has wheezing ever been severe enough to limit your child's speech to only one or two words at a time between breaths?"

The composite variable "allergy score" was calculated as the sum of positive responses to the questions: In the past 12 months, "Has your child had a problem with sneezing or a runny or blocked nose when he/she did not have a cold or the flu?", "Has this nose problem been accompanied by itchy/watery eyes?" and "Has your child ever had eczema?"

Missing data approach

Variables with missing values were imputed using predictive mean matching via the aregImpute function from the HMisc R package.[27] This method was chosen because it is designed for missing binary, categorical, or continuous variables. The primary analysis was conducted using a single imputation. Three sensitivity analyses were conducted using a dataset comprised of: (a) a different run of the single imputation, (b) the subset of participants (n=272) not missing FEV_1/FVC at the baseline (Year 6) visit, and (c) the subset of participants (n=190) not missing FeNO at the baseline (Year 6) visit.

Variable coding/standardization

After imputation, ordinal variables were recoded using integer values. For right skewed quantitative data (i.e., FeNO), we applied a natural log transform prior to imputation. All binary, ordinal, and continuous variables were standardized (i.e., calculated as (X-mean(X))/SD(X) for variable X).

Clustering algorithms

We performed clustering using the three most common methods identified in our literature review: hierarchical clustering, k-medoids, and latent class analysis (LCA). Hierarchical clustering constructs a hierarchy of clusters of participants using a measure of "dissimilarity" between pairs and observations.[28] The k-medoids method partitions participants into k clusters using algorithms which aim to minimize the dissimilarity between observations in a cluster and the center (medoid) of the cluster.[29] While k-means is closely related and a commonly applied method for continuous variables, k-medoids is an appropriate alternative for mixed variable types and is also more robust to outliers than k-means. LCA is a likelihood-based approach which identifies k latent classes (clusters) of

participants, with the latent classes driving the observed correlations in variables so that within each class the variables are independent (i.e., conditionally independent).[30] Guided by criteria for optimal number of clusters, we selected the same number of clusters (k) for all methods (see Supplement for details).

Describing cluster characteristics

To characterize the clusters, we first summarized the input variables by cluster using: (a) graphical heatmaps of input variable means[31,32] and (b) tabulated numerical summaries. We tested for differences across clusters using analysis of variance (ANOVA) for quantitative variables and chi-square/Fisher exact tests for categorical variables and identified simple rules for assigning participants to clusters informed by classification tree methods (see Supplement for further details).

Predicting asthma control after two years based on cluster assignment

To evaluate the temporal validity of the clusters,[33] we tested for differences in asthma endpoints (variables that served as proxies of asthma control, as best estimated using available variables in the CHS dataset) at approximately two years follow-up (Year 8) by cluster membership determined at Year 6. We considered the following asthma endpoints: asthma control (four levels described previously), rescue medication usage in the past 12 months (none, moderate: < 2 days/week, or frequent: 2+ days/week), percent predicted FEV_1, FEV_1/FVC, FeNO, wheeze ("wheezing or whistling in the chest in the past 12 months"), persistent cough (during the past 12 months—apart from a cold or chest infection —a dry cough at night that lasted >3 weeks or a cough first thing in the morning or at other times of the day lasting for as much as 3 months in a row), and chronic bronchitic symptoms (bronchitis, or report of a daily cough for 3 months in a row, or congestion or phlegm other than when accompanied by a cold).[34] We tested for differences across clusters using ANOVA or chi-square/Fisher exact tests. All analyses were conducted in R version 3.5.1 (http://www.R-project.org).

Results

Variable selection and reduction

We initially selected 15 variables of clinical relevance from the set of available CHS variables, based on the literature. After assessing correlations, we eliminated: baseline eczema (correlated with allergy symptom score, as expected) as well as frequency of rescue medication use and wheeze on exercise (both correlated with asthma control). Our final set of 12 variables (Table 1) included: sex, race/ethnicity (Hispanic, Non-Hispanic White, or Other), asthma control, FEV_1 percent predicted, FEV_1/FVC, log FeNO, categorized sex-specific body mass index (BMI) percentiles—healthy/underweight, overweight, obese— from Centers for Disease Control and Prevention growth charts,[35] allergy score, premature birth (4 weeks early)[36], age at asthma diagnosis, residential second hand exposure to tobacco smoke, and parental asthma.

As shown in Table 2, the 598 CHS participants with asthma were more than half male (58.2%), predominantly Hispanic (50.5%) or non-Hispanic White (36.1%), and had

relatively mild asthma (57% reported 0 wheezing attacks in the past 12 months). Most variables were relatively complete, missing for <10% of participants (see Supplement for details).

Selection criteria for the optimal number of clusters guided our choice of k=4 clusters for each of the three clustering methods (hierarchical clustering, k-medoids, and LCA). LCA identified the most clinically relevant clusters, with large differences across the four LCA clusters in: FeNO, FEV_1/FVC, percent predicted FEV_1, asthma control, and allergy score (all p<0.001, as shown in the Figure 2 heatmap). We used a simplified classification tree to group a new study population into the four LCA clusters using only FEV_1/FVC, FeNO, and percent predicted FEV_1 with an 81.7% classification accuracy. If we alter cut points to more clinically representative FEV_1/FVC, FEV_1% predicted, and FeNO values, the classification accuracy drops to 73.3% (see Figure 3).

The clusters identified by hierarchical clustering and k-medoids were similar to each other (Figure E3, Table E3) with strong differences across clusters by sex and race/ethnicity but less clinically measurable parameters and predictive ability over time (Figure 2, Figure E4).

Differences in outcomes at 2-year follow up

Two years after the initial clustering, there was evidence for differences across LCA clusters in asthma control (p=0.007), frequency of rescue medication use (p=0.038), percent predicted FEV_1 (p<0.001), FeNO (p<0.001), and wheeze over 12 months (p=0.049), with LCA Cluster 4 typically having the most poorly controlled asthma (Table 3). Hierarchical clustering and k-medoids clusters were different at follow-up for only: percent predicted FEV_1 (both p<0.001), FEV_1/FVC (p=0.071 and p<0.001, respectively), and FeNO (p=0.04 and p=0.252, respectively).

Sensitivity analyses

In all three sensitivity analyses for missing data, LCA produced Year 6 visit clusters with FeNO and FEV_1/FVC being key variables distinguishing clusters (p<0.001 for all sensitivity analyses, Figure E5a–c) and with cluster membership still predictive of differences in asthma control, rescue medication use, percent predicted FEV_1, FeNO, and FEV_1/FVC after two years of follow-up (Table E5a–c).

Discussion

Clustering has been a promising tool in asthma research to better understand asthma heterogeneity. Previous asthma clustering studies have suggested clinically interesting phenotypes with some similarities across studies. However, results have varied potentially partially due to the lack of standardized data analysis approaches. In this paper, we identified a framework of data analysis practices toward reproducibility including: addressing missing data, considering the distribution of input variables, variable selection, and evaluating a range of clustering methods. When applying these methods to schoolchildren with asthma in the CHS, we found the clusters using LCA—distinguished by FeNO, FEV_1/FVC, and percent predicted FEV_1—were predictive of control after two years of follow-up.

Our approach of using previous literature to inform a set of practice principles for asthma phenotyping studies was inspired by approaches in other fields, such as the checklist for prediction model development produced by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) group.[37] This and other efforts to ensure reproducibility of multiscale models[38] are supported by efforts of the NIH to promote rigor and reproducibility of research studies.[39] Previously, Prosperi et al[40] rigorously demonstrated how asthma clustering results in a given dataset can be sensitive to variations in methodology, including variable selection and variable transformation. In addition, Deliu et al[15] summarized common asthma clustering methodologies and identified key methodological challenges: inclusion of mixed variable types, sensitivity of results to the set of variables included, sensitivity of results to the clustering method, and the anticipated differences in phenotypes across different study populations.

In our clustering analysis in the CHS, we found the most clinically predictive clusters over time using LCA. Howard et al[41] reviewed LCA methodology and summarized the contributions of LCA-based analyses for asthma and wheeze phenotypes in children. It is well-recognized that there is no universally "best" clustering algorithm; different clustering methods produce different partitions on the same data, so in practice one needs to evaluate results from several algorithms to identify the most relevant algorithm for a given application.[42] In our data, the clusters identified by hierarchical clustering and k-medoids were less differentiated, and likely more similar to each other because both methods used the Gower distance matrix. This highlights the need to evaluate a wide range of clustering methods using different underlying methodology.

Our LCA clusters were distinguished by FeNO, FEV_1/FVC, and percent predicted FEV_1. Several, but not all[43] studies, have found FeNO to be a key feature of cluster membership in more symptomatic and allergic patients.[8,12,44–46] Mahut et al[43] found that FeNO did not help identify clinically useful asthma phenotypes but the analysis (using principal component analysis and k-means) was conducted in a small dataset (n=169). But a recent meta-analysis suggested that using FeNO to adjust asthma medication resulted in a lower number of exacerbations[47] and the 2020 NIH National Asthma Education and Prevention Program asthma guideline update suggested to consider incorporating FeNO into ongoing asthma management strategies.[48] While many specialists have access to FeNO equipment, not all practitioners can incorporate this biomarker into their asthma management plan. In contrast, other indicators such as parental asthma and asthma symptoms are more practical to measure. Identifying patients that belong to groups that have a higher risk of uncontrolled asthma can help identify them to be followed more frequently in clinic to ensure they are on an optimal controller regimen and are addressing co-morbidities and triggers.

Our CHS data analysis had strengths and limitations. The CHS is a large, prospective population-based cohort of schoolchildren in southern California. We chose the dataset because it is reflective of the local population of schoolchildren with asthma, but this group generally contains more mild cases and may be more homogenous than a study population from an asthma clinic or targeted asthma cohort. As CHS participants were followed longitudinally, we were able to temporally validate the clusters by evaluating how well cluster membership at Year 6 predicted asthma control outcomes at Year 8.

A limitation of our study included the *a priori* selection input variables for the clustering algorithms based on our review of the literature (Table 1) and that we were limited to variables available in the pre-existing CHS data resource. Given the hypothesis-generating goals of clustering-based phenotypes, we may have missed an unknown feature that contributes to heterogeneity in childhood asthma such as sociodemographic factors. The CHS is a population-based cohort study of schoolchildren, with participants data collected by questionnaire and field team visits to schools, so the CHS data did not include some clinical measures used in other clustering studies (e.g., airways hyperreactivity/ bronchodilator response, total IgE, allergic sensitization, eosinophilia, etc.) and overall more mild asthma. The questionnaire assessment of asthma symptoms and medication use in the CHS was for epidemiological research purposes and differed slightly from the methods indicated per the National Institutes of Health (NIH) or Global Initiative for Asthma (GINA) guidelines to measure asthma severity and control. The CHS did not contain direct clinical measures of atopy, however this may have been at least partially represented by FeNO, which was relevant in the LCA clusters and may indicate the importance of allergic inflammation in asthma phenotypes. Another limitation was the considerable number of participants missing data on variables assessed at in-person visits (FeNO, spirometry measures, and BMI). This missing data was by design, due to practical constraints on school-based visits in Year 6 of the cohort. Single imputation was used to fill in the missing values (imputation models included previous year FeNO, a strong predictor of current FeNO and previous year BMI, also a strong predictor of current BMI). In three additional sensitivity analyses, we found our results to be robust to different approaches to these missing values. This demonstrates the robustness of the LCA cluster results to real-world missing data, as will be encountered in future clustering studies using EHR.

Conclusion:

A literature-informed clustering approach applied to the Children's Health Study demonstrated latent class analysis clusters—represented by a classification tree containing only exhaled nitric oxide and spirometry measures—had long-term clinical relevance. Future work might adopt our approach for asthma clustering to the analysis of broader patient characteristics directly from real-world (i.e., electronic health record) data. This may contribute to understanding underlying biological mechanisms and develop personalized treatment and management plans for each phenotype.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

References

1. Centers for Disease Control, National Center for Health Statistics. FastStats Asthma. 2015; https:// www.cdc.gov/nchs/fastats/asthma.htm. Accessed March 15, 2021.

2. Centers for Disease Control. Uncontrolled Asthma Among Persons with Current Asthma. September 2014; https://www.cdc.gov/asthma/asthma_stats/uncontrolled_asthma.pdf. Accessed March 15, 2021.

3. Barnes PJ, Jonsson B, Klim JB. The costs of asthma. Eur Respir J. 1996;9(4):636–642. [PubMed: 8726924]

4. Asher I, Pearce N. Global burden of asthma among children. Int J Tuberc Lung Dis. 2014;18(11):1269–1278. [PubMed: 25299857]

5. Wenzel SE, Schwartz LB, Langmack EL, et al. Evidence that severe asthma can be divided pathologically into two inflammatory subtypes with distinct physiologic and clinical characteristics. Am J Respir Crit Care Med. 1999;160(3):1001–1008. [PubMed: 10471631]

6. Payne D, Bush A. Phenotype-specific treatment of difficult asthma in children. Paediatric respiratory reviews. 2004;5(2):116–123. [PubMed: 15135121]

7. National Asthma E, Prevention P. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. J Allergy Clin Immunol. 2007;120(5 Suppl):S94–138. [PubMed: 17983880]

8. Haldar P, Pavord ID, Shaw DE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med. 2008;178(3):218–224. PMC3992366 [PubMed: 18480428]

9. Moore WC, Meyers DA, Wenzel SE, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med. 2010;181(4):315–323. PMC2822971 [PubMed: 19892860]

10. Schatz M, Hsu JW, Zeiger RS, et al. Phenotypes determined by cluster analysis in severe or difficult-to-treat asthma. J Allergy Clin Immunol. 2014;133(6):1549–1556. [PubMed: 24315502]

11. Just J, Gouvis-Echraghi R, Couderc R, Guillemot-Lambert N, Saint-Pierre P. Novel severe wheezy young children phenotypes: boys atopic multiple-trigger and girls nonatopic uncontrolled wheeze. The Journal of allergy and clinical immunology. 2012;130(1):103–110.e108.

12. Fitzpatrick AM, Teague WG, Meyers DA, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. J Allergy Clin Immunol. 2011;127(2):382–389 e381–313. PMC3060668 [PubMed: 21195471]

13. Ross MK, Yoon J, van der Schaar A, van der Schaar M. Discovering Pediatric Asthma Phenotypes on the Basis of Response to Controller Medication Using Machine Learning. Ann Am Thorac Soc. 2018;15(1):49–58. PMC5822415 [PubMed: 29048949]

14. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. Am J Respir Crit Care Med. 2013;188(11):1303–1312. [PubMed: 24180417]

15. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. Pulmonary Therapy. 2016;2(1):19–41. [PubMed: 27512723]

16. Peters JM, Avol E, Navidi W, et al. A study of twelve Southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity. Am J Respir Crit Care Med. 1999;159(3):760–767. [PubMed: 10051248]

17. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006–1012. [PubMed: 19631508]

18. rpart: Recursive Partitioning and Regression Trees [computer program]. 2015.

19. Fingleton J, Travers J, Williams M, et al. Treatment responsiveness of phenotypes of symptomatic airways obstruction in adults. Journal of Allergy and Clinical Immunology. 2015;136(3):601–609. [PubMed: 25746966]

20. Vogel MA, Wong AK. PFS clustering method. IEEE transactions on pattern analysis and machine intelligence. 1979(3):237–245. [PubMed: 21868854]

21. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(2):411–423.

22. McConnell R, Berhane K, Yao L, et al. Traffic, susceptibility, and childhood asthma. Environ Health Perspect. 2006;114(5):766–772. PMC1459934 [PubMed: 16675435]

23. Gauderman WJ, Urman R, Avol E, et al. Association of improved air quality with lung development in children. N Engl J Med. 2015;372(10):905–913. [PubMed: 25738666]

24. Knudson RJ, Lebowitz MD, Holberg CJ, Burrows B. Changes in the normal maximal expiratory flow-volume curve with growth and aging. American Review of Respiratory Disease. 1983;127(6):725–734. [PubMed: 6859656]

25. Linn WS, Rappaport EB, Berhane KT, et al. Extended exhaled nitric oxide analysis in field surveys of schoolchildren: a pilot test. Pediatr Pulmonol. 2009;44(10):1033–1042. [PubMed: 19774622]

26. Linn WS, Rappaport EB, Eckel SP, et al. Multiple-flow exhaled nitric oxide, allergy, and asthma in a population of older children. Pediatr Pulmonol. 2013;48(9):885–896. 3748140 [PubMed: 23687084]

27. Harrell FE Jr, Dupont MC. The Hmisc Package. R Package, version. 2018:3.5–1.

28. Nielsen F. Hierarchical clustering. Introduction to HPC with MPI for Data Science: Springer; 2016:195–211.

29. Jin X, Han J. K-Medoids Clustering. Springer US, Boston, MA; 2010.

30. Vermunt JK, Magidson J. Latent class cluster analysis. Applied latent class analysis. 2002;11(89–106):60.

31. Garcia-Aymerich J, Benet M, Saeys Y, et al. Phenotyping asthma, rhinitis and eczema in M e DALL population-based birth cohorts: an allergic comorbidity cluster. Allergy. 2015;70(8):973–984. [PubMed: 25932997]

32. Loza MJ, Djukanovic R, Chung KF, et al. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. Respiratory research. 2016;17(1):165. [PubMed: 27978840]

33. Jain AK. Data clustering: 50 years beyond k-means. Paper presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2008.

34. Berhane K, Chang C-C, McConnell R, et al. Association of changes in air quality with bronchitic symptoms in children in California, 1993–2012. JAMA. 2016;315(14):1491–1501. [PubMed: 27115265]

35. Centers for Disease Control. A SAS program for the CDC growth charts. 2009; http://www.cdc.gov/nccdphp/dnpao/growthcharts/resources/sas.htm. Accessed March 15, 2021..

36. He H, Butz A, Keet CA, et al. Preterm Birth with Childhood Asthma: The Role of Degree of Prematurity and Asthma Definitions. Am J Respir Crit Care Med. 2015;192(4):520–523. PMC4595670 [PubMed: 26278798]

37. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. British Journal of Surgery. 2015;102(3):148–158. [PubMed: 25627261]

38. Altman E, Avrachenkov K, Ramanath S. Multiscale fairness and its application to resource allocation in wireless networks. Computer Communications. 2012;35(7):820–828.

39. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014;505(7485):612–613. PMC4058759 [PubMed: 24482835]

40. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. American journal of respiratory and critical care medicine. 2013;188(11):1303–1312. [PubMed: 24180417]

41. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing asthma phenotypes using machine learning approaches. Current allergy and asthma reports. 2015;15(7):38. [PubMed: 26143394]

42. Jain AK. Data clustering: 50 years beyond K-means. Pattern recognition letters. 2010;31(8):651–666.

43. Mahut B, Peyrard S, Delclaux C. Exhaled nitric oxide and clinical phenotypes of childhood asthma. Respir Res. 2011;12:65. PMC3126727 [PubMed: 21599913]

44. Hinks TS, Brown T, Lau LC, et al. Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3-like protein 1. J Allergy Clin Immunol. 2016;138(1):61–75. PMC4929135 [PubMed: 26851968]

45. Musk AW, Knuiman M, Hunter M, et al. Patterns of airway disease and the clinical diagnosis of asthma in the Busselton population. Eur Respir J. 2011;38(5):1053–1059. [PubMed: 21565919]

46. Couto M, Stang J, Horta L, et al. Two distinct phenotypes of asthma in elite athletes identified by latent class analysis. J Asthma. 2015;52(9):897–904. [PubMed: 26377281]

47. Petsky HL, Cates CJ, Kew KM, Chang AB. Tailoring asthma treatment on eosinophilic markers (exhaled nitric oxide or sputum eosinophils): a systematic review and meta-analysis. Thorax. 2018;73(12):1110–1119. [PubMed: 29858277]

48. Expert Panel Working Group of the National Heart L, Blood Institute a, coordinated National Asthma E, et al. 2020 Focused Updates to the Asthma Management Guidelines: A Report from the National Asthma Education and Prevention Program Coordinating Committee Expert Panel Working Group. J Allergy Clin Immunol. 2020;146(6):1217–1270. [PubMed: 33280709]

49. Spycher BD, Silverman M, Brooke AM, Minder CE, Kuehni CE. Distinguishing phenotypes of childhood wheeze and cough using latent class analysis. Eur Respir J. 2008;31(5):974–981. [PubMed: 18216047]

50. Agache I, Ciobanu C. Risk factors and asthma phenotypes in children and adults with seasonal allergic rhinitis. Phys Sportsmed. 2010;38(4):81–86. [PubMed: 21150146]

51. Benton AS, Wang Z, Lerner J, et al. Overcoming heterogeneity in pediatric asthma: tobacco smoke and asthma characteristics within phenotypic clusters in an African American cohort. J Asthma. 2010;47(7):728–734. PMC3325290 [PubMed: 20684733]

52. Just J, Gouvis-Echraghi R, Rouve S, et al. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. Eur Respir J. 2012;40(1):55–60. [PubMed: 22267763]

53. Just J, Saint-Pierre P, Gouvis-Echraghi R, et al. Wheeze phenotypes in young children have different courses during the preschool period. Ann Allergy Asthma Immunol. 2013;111(4):256–261 e251.

54. Lavoie-Charland É, Bérubé J-C, Laviolette M, Boulet L-P, Bossé Y. Multivariate asthma phenotypes in adults: the Quebec City case-control asthma cohort. Open Journal of Respiratory Diseases. 2013;3(04):133.

55. 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(5):1058–1067. [PubMed: 24075230]

56. Lemiere C SNG, Sava F, et al. Occupational asthma phenotypes identified by increased fractional exhaled nitric oxide after exposure to causal agents. J Allergy Clin Immunol. 2014;134(5):1063–1067. [PubMed: 25262466]

57. Moore WC, Hastie AT, Li X, et al. Sputum neutrophil counts are associated with more severe asthma phenotypes using cluster analysis. J Allergy Clin Immunol. 2014;133(6):1557–1563 e1555. PMC4040309

58. Sakagami T, Hasegawa T, Koya T, et al. Cluster analysis identifies characteristic phenotypes of asthma with accelerated lung function decline. J Asthma. 2014;51(2):113–118. [PubMed: 24102534]

59. Siroux V, Gonzalez JR, Bouzigon E, et al. Genetic heterogeneity of asthma phenotypes identified by a clustering approach. Eur Respir J. 2014;43(2):439–452. [PubMed: 24311777]

60. Wu W, Bleecker E, Moore W, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. J Allergy Clin Immunol. 2014;133(5):1280–1288. PMC4038417 [PubMed: 24589344]

61. Loureiro CC, Sa-Couto P, Todo-Bom A, Bousquet J. Cluster analysis in phenotyping a Portuguese population. Rev Port Pneumol (2006). 2015.

62. Mastalerz L, Celejewska-Wojcik N, Wojcik K, et al. Induced sputum supernatant bioactive lipid mediators can identify subtypes of asthma. Clin Exp Allergy. 2015;45(12):1779–1789. [PubMed: 26449970]

63. Park HW, Song WJ, Kim SH, et al. Classification and implementation of asthma phenotypes in elderly patients. Ann Allergy Asthma Immunol. 2015;114(1):18–22. [PubMed: 25455518]

64. Serrano-Pariente J, Rodrigo G, Fiz JA, et al. Identification and characterization of near-fatal asthma phenotypes by cluster analysis. Allergy. 2015;70(9):1139–1147. [PubMed: 26011771]

65. Sbihi H, Koehoorn M, Tamburic L, Brauer M. Asthma Trajectories in a Population-based Birth Cohort. Impacts of Air Pollution and Greenness. Am J Respir Crit Care Med. 2017;195(5):607–613. [PubMed: 27606967]

66. Sutherland ER, Goleva E, King TS, et al. Cluster analysis of obesity and asthma phenotypes. PLoS One. 2012;7(5):e36631. PMC3350517

67. Wu ST, Sohn S, Ravikumar KE, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Ann Allergy Asthma Immunol. 2013;111(5):364–369. PMC3839107 [PubMed: 24125142]

68. Jo KW, Ra SW, Chae EJ, et al. Three phenotypes of obstructive lung disease in the elderly. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease. 2010;14(11):1481–1488.

69. Fingleton J, Travers J, Williams M, et al. Treatment responsiveness of phenotypes of symptomatic airways obstruction in adults. J Allergy Clin Immunol. 2015;136(3):601–609. [PubMed: 25746966]

70. Williams-DeVane CR, Reif DM, Hubal EC, et al. Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes. BMC Syst Biol. 2013;7:119. PMC4228284 [PubMed: 24188919]

71. Meyer N, Dallinga JW, Nuss SJ, et al. Defining adult asthma endotypes by clinical features and patterns of volatile organic compounds in exhaled air. Respiratory research. 2014;15:136. [PubMed: 25431084]

72. Konstantellou E, Papaioannou AI, Loukides S, et al. Persistent airflow obstruction in patients with asthma: Characteristics of a distinct clinical phenotype. Respir Med. 2015;109(11):1404–1409. [PubMed: 26412805]

73. Ortega H, Prazma C, Suruki RY, Li H, Anderson WH. Association of CHI3L1 in African-Americans with prior history of asthma exacerbations and stress. J Asthma. 2013;50(1):7–13. [PubMed: 23190377]

74. Garcia-Aymerich J, Benet M, Saeys Y, et al. Phenotyping asthma, rhinitis and eczema in MeDALL population-based birth cohorts: an allergic comorbidity cluster. Allergy. 2015;70(8):973–984. [PubMed: 25932997]

75. Boudier A, Curjuric I, Basagana X, et al. Ten-year follow-up of cluster-based asthma phenotypes in adults. A pooled analysis of three cohorts. Am J Respir Crit Care Med. 2013;188(5):550–560. [PubMed: 23777340]

76. Ranciere F, Nikasinovic L, Bousquet J, Momas I. Onset and persistence of respiratory/allergic symptoms in preschoolers: new insights from the PARIS birth cohort. Allergy. 2013;68(9):1158–1167. [PubMed: 23919292]

77. Weinmayr G, Keller F, Kleiner A, et al. Asthma phenotypes identified by latent class analysis in the ISAAC phase II Spain study. Clin Exp Allergy. 2013;43(2):223–232. [PubMed: 23331563]

78. Siroux V, Basagana X, Boudier A, et al. Identifying adult asthma phenotypes using a clustering approach. Eur Respir J. 2011;38(2):310–317. [PubMed: 21233270]

79. Kim TB, Jang AS, Kwon HS, et al. Identification of asthma clusters in two independent Korean adult asthma cohorts. Eur Respir J. 2013;41(6):1308–1314. [PubMed: 23060627]

80. Rootmensen G, van Keimpema A, Zwinderman A, Sterk P. Clinical phenotypes of obstructive airway diseases in an outpatient population. J Asthma. 2016;53(10):1026–1032. [PubMed: 27366830]

81. Yamada H, Masuko H, Yatagai Y, et al. Role of Lung Function Genes in the Development of Asthma. PLoS One. 2016;11(1):e0145832. PMC4709100

82. Zaihra T, Walsh CJ, Ahmed S, et al. Phenotyping of difficult asthma using longitudinal physiological and biomarker measurements reveals significant differences in stability between clusters. BMC Pulm Med. 2016;16(1):74. PMC4862112 [PubMed: 27165150]

83. Weatherall M, Travers J, Shirtcliffe PM, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis. Eur Respir J. 2009;34(4):812–818. [PubMed: 19357143]

84. Wu W, Bang S, Bleecker ER, et al. Multiview Cluster Analysis Identifies Variable Corticosteroid Response Phenotypes in Severe Asthma. Am J Respir Crit Care Med. 2019;199(11):1358–1367. PMC6543720 [PubMed: 30682261]

85. Kaneko Y, Masuko H, Sakamoto T, et al. Asthma phenotypes in Japanese adults - their associations with the CCL5 and ADRB2 genotypes. Allergol Int. 2013;62(1):113–121.

86. Amelink M, de Nijs SB, de Groot JC, et al. Three phenotypes of adult-onset asthma. Allergy. 2013;68(5):674–680. [PubMed: 23590217]

87. Howrylak JA, Fuhlbrigge AL, Strunk RC, et al. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. J Allergy Clin Immunol. 2014;133(5):1289–1300, 1300 e1281–1212. PMC4047642 [PubMed: 24892144]

88. Ortega H, Li H, Suruki R, et al. Cluster analysis and characterization of response to mepolizumab. A step closer to personalized medicine for patients with severe asthma. Ann Am Thorac Soc. 2014;11(7):1011–1017. [PubMed: 24983709]

89. Newby C, Heaney LG, Menzies-Gow A, et al. Statistical cluster analysis of the British Thoracic Society Severe refractory Asthma Registry: clinical outcomes and phenotype stability. PLoS One. 2014;9(7):e102987. PMC4109965

90. Loza MJ, Adcock I, Auffray C, et al. Longitudinally Stable, Clinically Defined Clusters of Patients with Asthma Independently Identified in the ADEPT and U-BIOPRED Asthma Studies. Ann Am Thorac Soc. 2016;13 Suppl 1:S102–103.

91. Zoratti EM, Krouse RZ, Babineau DC, et al. Asthma phenotypes in inner-city children. J Allergy Clin Immunol. 2016;138(4):1016–1029. PMC5104222 [PubMed: 27720016]

92. Baptist AP, Hao W, Karamched KR, et al. Distinct Asthma Phenotypes Among Older Adults with Asthma. J Allergy Clin Immunol Pract. 2018;6(1):244–249 e242. PMC5897052 [PubMed: 28757370]

93. Ilmarinen P, Tuomisto LE, Niemela O, et al. Cluster Analysis on Longitudinal Data of Patients with Adult-Onset Asthma. J Allergy Clin Immunol Pract. 2017;5(4):967–978 e963. [PubMed: 28389304]

94. Sendin-Hernandez MP, Avila-Zarza C, Sanz C, et al. Cluster Analysis Identifies 3 Phenotypes within Allergic Asthma. J Allergy Clin Immunol Pract. 2018;6(3):955–961 e951.

95. Wang L, Liang R, Zhou T, et al. Identification and validation of asthma phenotypes in Chinese population using cluster analysis. Ann Allergy Asthma Immunol. 2017;119(4):324–332. [PubMed: 28866310]

96. Panico L, Stuart B, Bartley M, Kelly Y. Asthma trajectories in early childhood: identifying modifiable factors. PLoS One. 2014;9(11):e111922. PMC4224405

97. Pite H, Gaspar A, Morais-Almeida M. Preschool-age wheezing phenotypes and asthma persistence in adolescents. Allergy Asthma Proc. 2016;37(3):231–241. [PubMed: 27074977]

98. Schoos AM, Chawes BL, Rasmussen MA, et al. Atopic endotype in childhood. J Allergy Clin Immunol. 2016;137(3):844–851 e844.

99. Westman M, Kull I, Lind T, et al. The link between parental allergy and offspring allergic and nonallergic rhinitis. Allergy. 2013;68(12):1571–1578. [PubMed: 24117663]

100. Sonnenschein-van der Voort AM, Arends LR, de Jongste JC, et al. Preterm birth, infant weight gain, and childhood asthma risk: a meta-analysis of 147,000 European children. J Allergy Clin Immunol. 2014;133(5):1317–1329. PMC4024198 [PubMed: 24529685]

Identification

Records identified through database search (n = 1,091)

Screening

Records screened (n = 1,091)

Records excluded (title, abstract, language, subject) (n = 991)

Eligibility

Full-text articles assessed for eligibility (n = 100)

Full-text articles excluded due to clinical content (n = 23)

Included
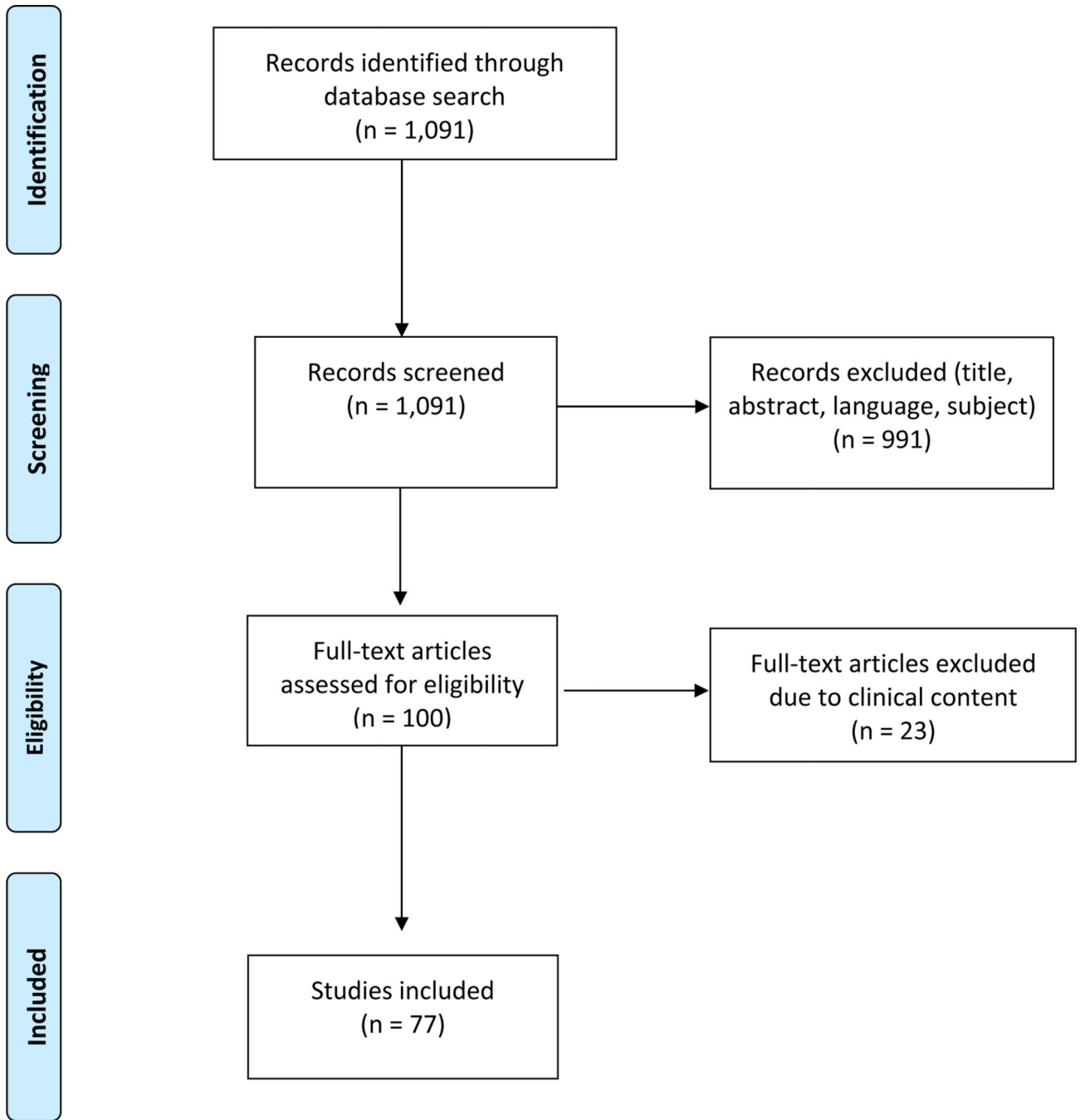
Studies included (n = 77)

Figure 1.
Article selection for the literature review, displayed using a PRISMA flow diagram
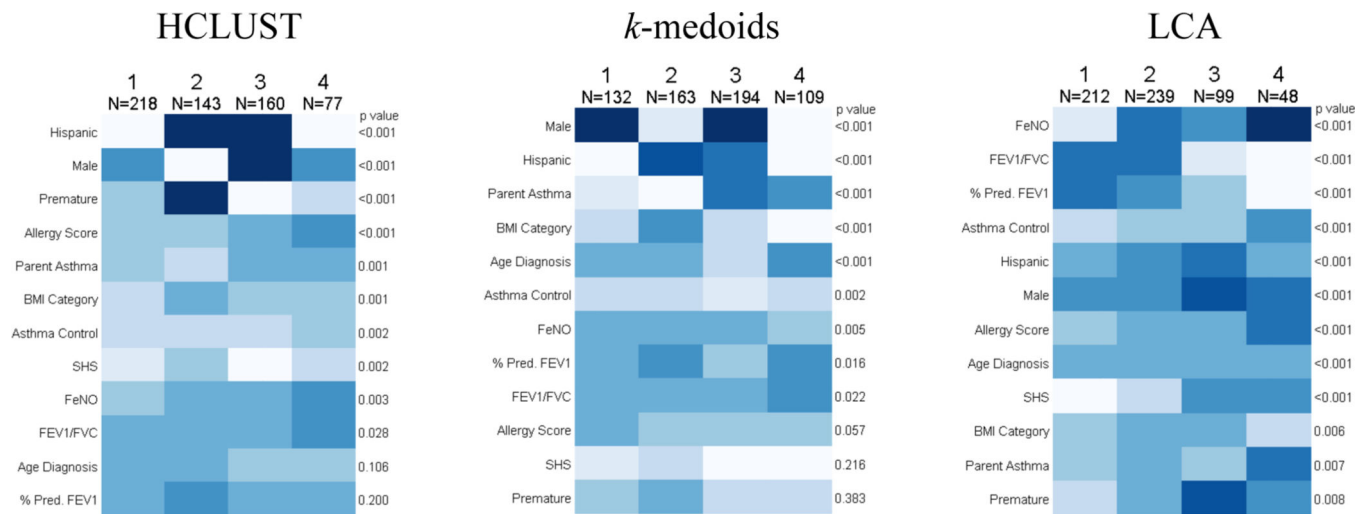
Figure 2.
Characterization of the clusters resulting from each of three clustering methods (HCLUST: hierarchal clustering, *k*-medoids, and LCA: latent class analysis), using heatmaps to represent the cluster-specific mean values* of input variables.[†] Darker values indicate larger means. Input variables (rows) are ordered by p-values for differences across clusters. Clusters (columns) are labeled with the number of participants in that cluster.
* Complementary numerical summaries are presented in Table E4.
[†] Input variables: Male is a binary indicator for sex (male). For simple visual presentation, here we use a binary indicator of Hispanic rather than the nominal 3-level race/ethnicity variable (Hispanic, non-Hispanic White, Other) input to the clustering algorithms. BMI Category is the BMI percentile category, SHS is secondhand smoke, and % pred FEV1 is percent predicted FEV_1.
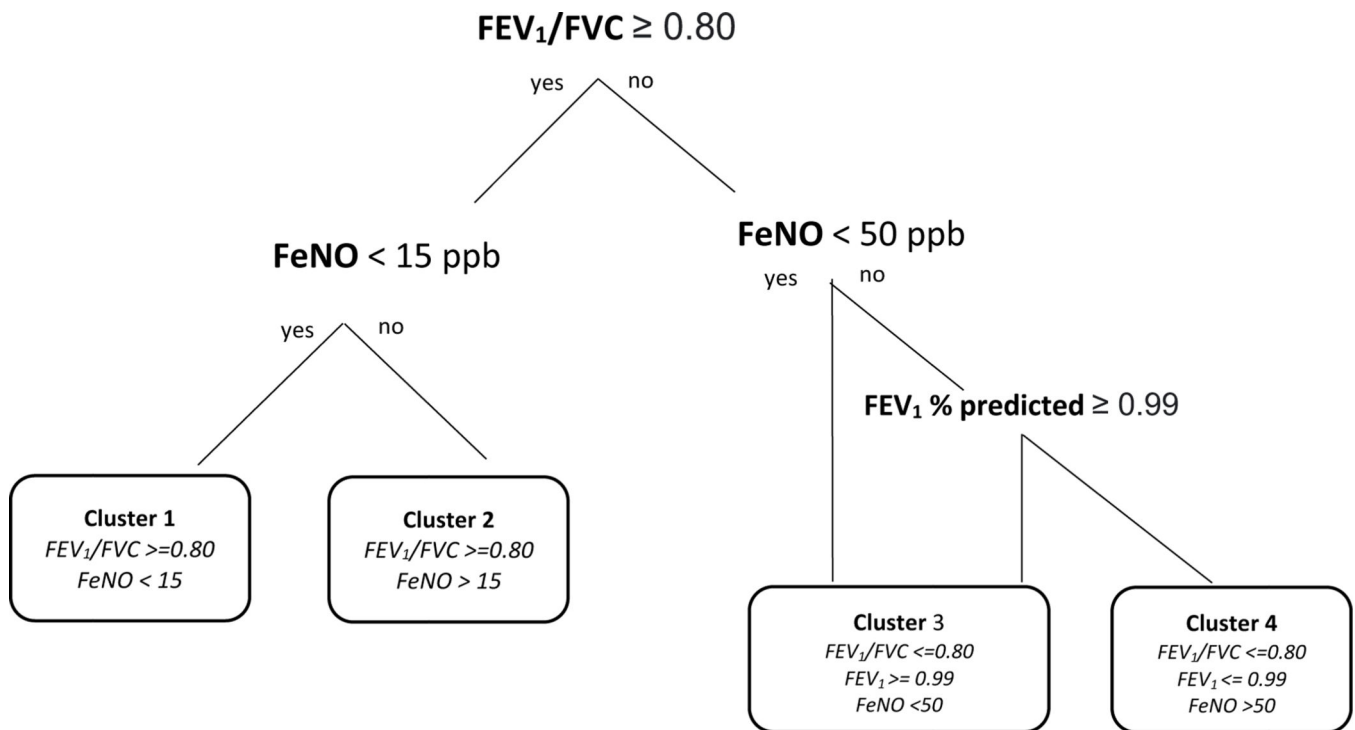
Figure 3.
Representation of the four clusters identified in the Children's Health Study (CHS) by Latent Class Analysis (LCA), using a flow chart with decision points informed by a simplified classification (decision) tree. This simple flow chart did not exactly reproduce the clusters, but did yield an 81.7% classification accuracy within the 30% holdout test dataset.*

* Modification of this flow chart to use decision points based on the clinically relevant values of $FEV_1/FVC >= 85$, FeNO <25, and FEV_1 80% predicted yielded a classification accuracy of 73.3% in the test dataset.

Table 1.

The final 12 Children Health Study (CHS) variables used in the clustering algorithms, supported by literature to guide their selection

Variable in the CHS	Related concepts in previous literature
Demographics	**Demographics**
Sex (male, female)	Sex[8–10,43–45,49–65]
Race/Ethnicity (Hispanic, Non-Hispanic White, Asian, African American, Other)	Race/Ethnicity[9,10,12,57,60,65,66]
Asthma Control	**Asthma Control**
Asthma Control in last 12 months 0: no wheeze attacks; 1: 1–3 attacks, no shortness of breath, no sleep disturbance; 2: 1–3 attacks, shortness of breath, and/or sleep disturbance; 3: 4+ attacks Composite derived from questions: In the past 12 months, • How many attacks of wheezing has your child had? • How often, on average, has your child's sleep been disturbed due to wheezing? • Has wheezing ever been severe enough to limit your child's speech to only one or two words at a time between breaths?	Severity/Control classification: mild, mod or severe;[9,53,62,64,67] risk composite score[61] Symptom questionnaires: ACT,[44,51,61] SOA,[61] AEQ,[66] ALQ,[61] SGRQ,[68,69] GHQ,[64] Juniper[8,61,70,71] Medication: type of controller,[64,71,72] number of controllers,[12,57,66] oral steroid dose and days,[43,57] composite score[9,61] Symptoms: days/week,[12,43] exacerbations/year,[59,63,73,74] type of symptoms[74–77]
Spirometry/Lung Function	**Spirometry/Lung Function**
FEV$_1$, % predicted	FEV$_1$-related: FEV$_1$ % predicted [9,12,43–45,51,52,54,57,61,62,66,68,69,73,75,78–82], FEV$_1$[60,70,71], FEV$_{0.5}$[49,55], FEV$_1$% change post-bronchodilator[9,12,63,78,80,83], FEV$_1$ change post-bronchodilator,[70] decline FEV$_1$[58], FEV$_1$ improvement at altitude[71], FEV$_1$ <80%[59]
FEV1/FVC ratio (%)	FEV1/FVC[9,12,43–46,51,57,58,60,66,68,69,72,79,82–95]
Phenotypes	**Phenotypes**
FeNO, ppb (natural log transformed for analysis)	FeNO,[8,12,43–46,57,66,69,70] post-bronchodilator FeNO[56]
BMI percentile (categorized sex-specific) (<85: Healthy/underweight, 85 and < 95: Overweight, 95: Obese)	BMI percentile or z-score,[8,9,51,53,57,60,66] BMI value[10,43–45,52,54,61–63,69,70,73,79,80,82]
Allergy Score Sum of positive reports to the following 3 questions (range: 0–3): • Has your child ever had eczema? • In the past 12 mo, has your child had a problem with sneezing or a runny or blocked nose when he/she did not have a cold or the flu? • Has this nose problem been accompanied by itchy/watery eyes?	Allergic or atopic (Y/N),[8,10,11,44,46,52,53,60,61,64,70,74,78,96] skin prick test,[45,46,49,50,54–56,62–64,72,97,98] blood test,[11,12,46,53,59,64,70,75,80]
Age of asthma diagnosis/onset At what age was child first diagnosed with asthma by a doctor?	Age of asthma onset,[8–10,44,51,54,57–62,64,66,69,74,78,79,81,97] duration of asthma[12,52,57,58,61,63,73,78,82]
Associated Factors	**Associated Factors**
Secondhand smoke exposure Did anyone living in child's home currently smoke inside the home?	Smoker (self),[9,43,44,50,54,56,63,64,69,83] passive smoke exposure (Y/N)[9,10]
Parental asthma Has a doctor ever said that this child's biological mother/father had asthma?	Parental asthma[9,11,50,52,53,73,97,99]
Premature (4 weeks early; i.e., <37 weeks gestation) Was your child born prematurely? How many weeks early?	Prematurity[100]

ACT=Asthma Control Test; SOA=Severity of Asthma score; AEQ=Asthma Evaluation Questionnaire; ALQ=Asthma Life Quality; SGRQ=St. George's Respiratory Questionnaire; GHQ=General Health Questionnaire; FEV1=Forced expiratory volume in 1 second; FeNO=fractional ide; BMI=body mass index

Table 2.

Summary of characteristics of the n=598 southern California Children's Healthy Study (CHS) participants with asthma, at the Year 6 (2006–2007) visit.

Participant characteristic	N (%) or Mean (SD)
Age (SD)	11.3 (0.98)
Male, n (%)	348 (58.2)
Race/Ethnicity, n (%)	
Hispanic	302 (50.5)
Non-Hispanic White	216 (36.1)
Asian [*]	21 (3.5)
African American [*]	14 (2.3)
Other [*]	45 (7.5)
Asthma Control, n (%)	
0 attack	341 (57.0)
1–3 attack, no SOB, no sleep	99 (16.6)
1–3 attack, SOB or sleep	85 (14.2)
4+ attack	73 (12.2)
% Predicted FEV1; mean (SD)	106.4 (13.6)
Missing, n (%) [†]	323 (54.0)
FEV1/FVC; mean (SD)	0.85 (0.06)
Missing, n (%) [†]	326 (54.5)
FeNO, ppb; mean (SD) [‡]	21.3 (2.5)
Missing, n (%) [†]	408 (68.2)
Body Mass Index Percentile Category, n (%)	
Healthy/Underweight, <85%	315 (52.7)
Overweight, 85–94.9%	119 (19.9)
Obese, 95%	164 (27.4)
Missing, n (%) [†]	254 (42.5)
Allergy Score, n (%) [§]	
0	53 (8.9)
1	87 (14.5)
2	332 (55.5)
3	126 (21.1)
Premature, n (%)	46 (7.7)
Missing, n (%) [†]	22 (3.7)
Age of asthma diagnosis, n (%) [**]	

Participant characteristic	N (%) or Mean (SD)
4 years	354 (59.2)
5–9 years	208 (34.8)
10 years	36 (6.0)
Missing, n (%) [†]	10 (1.7)
Secondhand tobacco smoke exposure, n (%)	22 (3.7)
Missing, n (%) [†]	35 (5.9)
Parent with asthma, n (%)	245 (41.0)
Missing, n (%) [†]	53 (8.9)

[*] For the cluster analysis, Asian, African American and Other were collapsed into a new "Other" category (N=80, 13.4%) due to small sample sizes.

[†] Missing values were imputed in the cluster analysis, with imputation R^2 of: 0.99 for % predicted FEV_1, 0.98 for FEV1/FVC, 0.88 for log FeNO, 0.87 for BMI percentile.

[‡] Geometric mean and standard deviation (SD) are presented for right-skewed FeNO

[§] Sum of positive reports of runny nose without cold/flu, itchy/watery eye accompanying this runny nose, eczema

[**] For the cluster analysis we used the ordinal questionnaire responses: <1 year old, 1 year old, 2 years old, 3 years old, 4 years old, 5 years old, 6 years old, 7 years old, 8–9 years old, 10 or more years old

Table 3.

Summaries of asthma endpoints * after two years of follow-up, by clusters determined at the Year 6 (2006–2007) visit using three clustering methods (HCLUST: hierarchal clustering, k-medoids, and LCA: latent class analysis). P-values are for tests of differences across clusters.

	Overall	HCLUST					k-medoids					LCA				
		1	2	3	4	p-value	1	2	3	4	p-value	1	2	3	4	p-value
Asthma control, N (%)						0.438					0.764					0.007
0 attack	303 (60.4)	107 (58.5)	76 (64.4)	85 (63.0)	35 (53.0)		60 (56.6)	85 (63.4)	100 (61.3)	58 (58.6)		112 (65.1)	129 (62.3)	49 (57.0)	13 (35.1)	
1–3 attack, no SOB, no Sleep disturbance	78 (15.5)	36 (19.7)	11 (9.3)	21 (15.6)	10 (15.2)		20 (18.9)	14 (10.4)	29 (17.8)	15 (15.2)		31 (18.0)	28 (13.5)	14 (16.3)	5 (13.5)	
1–3 attack, SOB or sleep disturbance	68 (13.5)	23 (12.6)	18 (15.3)	16 (11.9)	11 (16.7)		14 (13.2)	19 (14.2)	21 (12.9)	14 (14.1)		16 (9.3)	27 (13.0)	13 (15.1)	12 (32.4)	
4+ attack	53 (10.6)	17 (9.3)	13 (11.0)	13 (9.6)	10 (15.2)		12 (11.3)	16 (11.9)	13 (8.0)	12 (12.1)		13 (7.6)	23 (11.1)	10 (11.6)	7 (18.9)	
Rescue medication use, N (%)						0.833					0.407					0.038
None	276 (55.6)	102 (57.3)	67 (56.8)	72 (53.7)	35 (53.0)		52 (50.0)	72 (54.1)	91 (55.8)	61 (63.5)		102 (60.7)	111 (54.1)	50 (58.1)	13 (35.1)	
Moderate	121 (24.4)	47 (26.4)	26 (22.0)	32 (23.9)	16 (24.2)		31 (29.8)	29 (21.8)	42 (25.8)	19 (19.8)		38 (22.6)	49 (23.9)	24 (27.9)	10 (27.0)	
Frequent	99 (20.0)	29 (16.3)	25 (21.2)	30 (22.4)	15 (22.7)		21 (20.2)	32 (24.1)	30 (18.4)	16 (16.7)		28 (16.7)	45 (22.0)	12 (14.0)	14 (37.8)	
% Predicted FEV1, mean (SD)	105.76 (14.12)	105.35 (13.87)	111.31 (11.75)	104.97 (14.34)	98.47 (14.77)	<0.001	99.65 (13.27)	110.10 (13.04)	103.33 (13.62)	110.33 (14.30)	<0.001	106.92 (14.02)	108.85 (13.35)	101.98 (11.58)	92.65 (15.58)	<0.001
FEV1/FVC	0.85 (0.06)	0.85 (0.07)	0.87 (0.05)	0.84 (0.06)	0.86 (0.08)	0.071	0.82 (0.08)	0.86 (0.06)	0.85 (0.06)	0.88 (0.06)	<0.001	0.86 (0.06)	0.87 (0.05)	0.81 (0.06)	0.79 (0.10)	<0.001
FeNO, ppb; mean (SD) †	24.5 (2.4)	19.1 (2.2)	27.4 (2.2)	29.1 (2.5)	28.8 (2.6)	0.04	30.3 (2.2)	24.5 (2.2)	24.5 (2.7)	20.1 (2.2)	0.252	13.5 (1.8)	31.2 (2.2)	37.7 (2.2)	64.1 (2.1)	<0.001
Wheeze past 12 mo, N (%)	212 (65.6)	82 (64.1)	43 (59.7)	54 (70.1)	33 (71.7)	0.439	51 (64.6)	52 (63.4)	66 (68.8)	43 (65.2)	0.887	59 (58.4)	89 (67.9)	37 (62.7)	27 (84.4)	0.049
Persistent cough, N (%)	83 (20.2)	32 (21.5)	17 (18.1)	21 (19.1)	13 (22.8)	0.865	23 (26.7)	22 (20.6)	23 (16.9)	15 (18.5)	0.342	25 (17.4)	35 (21.2)	15 (21.1)	8 (26.7)	0.651

	Overall	HCLUST					k-medoids					LCA				
		1	2	3	4	p-value	1	2	3	4	p-value	1	2	3	4	p-value
Chronic bronchitic symptoms, N (%)	161 (39.1)	32 (21.5)	17 (18.1)	21 (19.1)	13 (22.8)	0.916	38 (43.2)	46 (42.6)	45 (34.1)	32 (38.1)	0.458	52 (35.6)	69 (40.6)	24 (36.9)	16 (51.6)	0.381

*
At year 8, endpoints were missing for some of the N=598 participants clustered at year 6, with the following frequency: 96 for asthma severity, 102 for rescue medication use, 331 for % predicted FEV$_1$, 333 for FEV$_1$/FVC, 434 for FeNO, 275 for wheeze, 188 for persistent cough, 186 for chronic bronchitic symptoms.

†
Geometric mean and standard deviation (SD) are presented for right-skewed FeNO.