Challenges in Identifying Asthma Subgroups Using Unsupervised Statistical Learning Techniques

Mattia C. F. Prosperi[1,2], Umit M. Sahiner[3], Danielle Belgrave[1,2], Cansin Sackesen[3], Iain E. Buchan[1], Angela Simpson[2], Tolga S. Yavuz[3], Omer Kalayci[3]*, and Adnan Custovic[2]*

[1]Centre for Health Informatics, Institute of Population Health, and [2]Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester, United Kingdom; and [3]Hacettepe University School of Medicine, Pediatric Allergy and Asthma Unit, Ankara, Turkey

Rationale: **Unsupervised statistical learning techniques, such as exploratory factor analysis (EFA) and hierarchical clustering (HC), have been used to identify asthma phenotypes, with partly consistent results. Some of the inconsistency is caused by the variable selection and demographic and clinical differences among study populations.**
Objectives: **To investigate the effects of the choice of statistical method and different preparations of data on the clustering results; and to relate these to disease severity.**
Methods: **Several variants of EFA and HC were applied and compared using various sets of variables and different encodings and transformations within a dataset of 383 children with asthma. Variables included lung function, inflammatory and allergy markers, family history, environmental exposures, and medications. Clusters and original variables were related to asthma severity (logistic regression and Bayesian network analysis).**
Measurements and Main Results: **EFA identified five components (eigenvalues $\geq$ 1) explaining 35% of the overall variance. Variations of the HC (as linkage-distance functions) did not affect the cluster inference; however, using different variable encodings and transformations did. The derived clusters predicted asthma severity less than the original variables. Prognostic factors of severity were medication usage, current symptoms, lung function, paternal asthma, body mass index, and age of asthma onset. Bayesian networks indicated conditional dependence among variables.**
Conclusions: **The use of different unsupervised statistical learning methods and different variable sets and encodings can lead to multiple and inconsistent subgroupings of asthma, not necessarily correlated with severity. The search for asthma phenotypes needs more careful selection of markers, consistent across different study populations, and more cautious interpretation of results from unsupervised learning.**

Keywords: asthma; children; clustering; machine learning; endotypes

AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

Unsupervised statistical learning techniques have been used to identify latent subgroups of children and adults with asthma who display different patterns of clinical features. The results were only partly consistent across different studies, giving rise to different subgroupings.

What This Study Adds to the Field

The observed heterogeneity reflects differences in demographic and clinical characteristics of the populations examined. However, we also demonstrate that such inconsistencies may be an artifact of the clustering techniques used, and of the variable encodings and transformations (e.g., discretization and dimension reduction).

Despite efforts made by the pharmaceutical industry and academia, asthma remains poorly understood, with a modest drug armamentarium (1). There is increasing recognition that asthma is a heterogeneous disease with multiple disease variants, which may have a similar clinical presentation, but differ in their etiology and pathogenesis (2, 3). It is likely that these different asthma subgroups (sometimes referred to as asthma endotypes [3]) have different causative mechanisms, and may require different treatments. Appropriate identification of such asthma subgroups is a critically important first step toward understanding their specific underlying biologic mechanisms, which is a key building block for therapeutic target identification and the development of novel treatments (4). This is a prerequisite for the move toward personalized or stratified health care to optimize clinical management and prevention of asthma (5).

Computer-assisted reasoning can facilitate the exploration of rich clinical data sets to enable better understanding of disease subgroups and their pathophysiology, and optimization of existing treatments. A data-driven approach with unsupervised statistical learning techniques can be used for discovery of latent asthma phenotypes, which can be derived based on a series of observable disease manifestations, instead of using predetermined classifications proposed by committees of experts. Several previous studies applied principal components analysis, exploratory factor analysis (EFA), partitioning clustering, hierarchical clustering (HC), and other techniques to identify latent groups and associated symptom patterns among adults (6–8) and children (9–12) with asthma. The results have been inconsistent. This inconsistency may be explained in part by natural heterogeneity (differences in the demographic or clinical characteristics of the populations studied), and in part by artifacts of data

(*Received in original form April 12, 2013; accepted in final form October 23, 2013*)

* These authors contributed equally.

Supported in part by University of Manchester's Health Research Centre funded by the MRC grant MR/K006665/1.

Author Contributions: M.C.F.P., study concept, data analysis, and manuscript writing. U.M.S., patient recruitment and data acquisition. D.B., statistical analysis. C.S., patient recruitment and data acquisition. I.E.B., statistical/informatics review and manuscript review. A.S., study concept and manuscript review. T.S.Y., patient recruitment and data acquisition. O.K., study design, clinical review, and manuscript review. A.C., study design, study concept, and manuscript review.

Correspondence and requests for reprints should be addressed to Mattia C. F. Prosperi, M.Eng., Ph.D., Centre for Health Informatics, Institute of Population Health, University of Manchester, 1st Floor, Jean McFarlane Building, Room 1.314, Oxford Road, Manchester M13 9PL, UK. E-mail: mattia.prosperi@manchester.ac.uk

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

Am J Respir Crit Care Med Vol 188, Iss. 11, pp 1303–1312, Dec 1, 2013

Originally Published in Press as DOI: 10.1164/rccm.201304-0694OC on November 1, 2013
Internet address: www.atsjournals.org

processing and analysis. We hypothesize that the subgrouping of asthma from typical study datasets is influenced by investigators' choice of factors, encoding/categorization and transformation (including dimensionality reduction) of variables, and choice of statistical method. To investigate this hypothesis, we compared variations of HC and EFA, with respect to different encodings and subsets of symptoms, markers, and diagnoses studied in populations of children with asthma. We then related the clustering to physician-reported asthma severity, and also considered which of the original variables (apart from cluster memberships) best predicted severity.

METHODS

Study Population

Children aged 6–18 years were recruited from the Pediatric Asthma Clinic at the Hacettepe University, Ankara, Turkey. Parents were interviewed by a pediatrician using a modified ISAAC questionnaire (13) to ascertain information on symptoms and prescribed medications. Children completed skin tests, spirometry (14), and measurement of bronchodilator reversibility. Those with a negative reversibility test underwent either methacholine or exercise challenge test to measure airway hyperresponsiveness (15–17). Blood sample was collected for measurement of eosinophils and total serum IgE.

Asthma was defined as all three of the following: (*1*) physician-diagnosed asthma, (*2*) current use of asthma medication, and (*3*) either bronchodilator reversibility or airway hyperresponsiveness (positive methacholine or exercise challenge test). Asthma severity was categorized into three ordinal categories (mild, moderate, and severe) using Global Initiative for Asthma guidelines (http://www.ginasthma.org/), based on the clinical features present and the patient's current step of the medication regimen.

Variables Used

The following variables were used:

Asthma symptoms and exacerbations: Presence of asthma-related symptoms within the past 4 weeks, number of asthma exacerbations within the past year, and hospitalization for acute asthma (ever).

Interview-derived variables: Age, sex, age of asthma onset, physician-diagnosed allergic rhinitis, conjunctivitis, urticaria and/or eczema, family history of asthma, and presence of smokers and pets or animals in the home.

Objective measurements: Height, weight, body mass index (BMI; standardized for age and growth and sex) (18), serum eosinophil number or percentage, and total serum IgE.

Medication usage: Use of short-acting β_2-agonists (SABA); inhaled corticosteroids (ICS), dose expressed as beclometasone-equivalent; long-acting β_2-agonists (LABA); and leukotriene receptor antagonist.

Lung function: % predicted FEV_1, FVC, forced expiratory flow ($FEF_{25–75}$), and FEV_1/FVC ratio.

Bronchodilator reversibility: Greater than or equal to 12% increase in FEV_1 following administration of 200 μg of inhaled albuterol.

Airway hyperresponsiveness: Provocative concentration of methacholine causing a 20% decline in FEV_1 (PC_{20}) less than or equal to 8 mg/ml (16) or greater than or equal to 10% reduction in FEV_1 following exercise challenge (17, 19–21).

Atopic sensitization: Wheal 3 mm greater than negative control to at least one allergen.

Statistical Methods

Variables were encoded either as raw mixed types or categorized with equal-frequency binning and projected into binary dummy variables. No missing values were present, apart from the alternative measures of airway hyperresponsiveness. Variables with a relative frequency below 1% were excluded. We performed EFA both by means of multiple factor analysis and principal component analysis (22). To facilitate visualization of weightings on dimensions, variables were grouped together in terms of (*1*) lung function (% predicted FEV_1, % predicted FVC, FEV_1/FVC ratio, $FEF_{25–75}$, bronchodilator reversibility, airway hyperresponsiveness), (*2*) markers of severity or exacerbation (symptoms within the past 4 wk, number of attacks within the last year, hospitalization), (*3*) family history, (*4*) comorbidities and atopy (rhinitis, eczema, sensitization, % eosinophils, total IgE); (*5*) environmental factors (exposure to tobacco smoke, pet ownership), (*6*) asthma medication (ICS, LABA, montelukast, and any combination), and (*7*) general characteristics (age, sex, BMI, age of onset of wheeze).

We applied different HC methods (23, 24) to the dataset, varying distance measures, linkage functions, feature selection (22, 25), and then identifying an optimal partition of the inferred trees (26). For control, a set of random trees and clusters was also created. All different variations of HC were mutually compared in a so-called "meta" HC, using the adjusted Rand index (27) and the Penny-Hendy index (28) as measures of trees and clusters similarity. Classical

TABLE 1. CHARACTERISTICS OF THE STUDY POPULATION (N = 383)

Variables	Value
Categorical	
Sex, male	60.6%
BMI, obese or overweight	25.3%
Allergic rhinitis	43.3%
Allergic conjunctivitis	2.3%
Eczema	3.4%
Atopy	59.6%
Mother's	
Asthma	4.4%
Allergic rhinitis	8.3%
Eczema	1.8%
Urticaria	0.5%
Father's	
Asthma	6.0%
Allergic rhinitis	5.5%
Eczema	1.8%
Urticaria	0.3%
Exposure to tobacco smoke	36.0%
Pet ownership	8.1%
Medications	
Budesonide	32.9%
Fluticasone	12.8%
LABA	9.9%
Montelukast	5.7%
Any drug in addition to SABA	46.7%
Asthma symptoms within the last 4 wk	17.0%
⩾1 hospitalization for acute asthma exacerbation	8.3%
Asthma severity	
Mild	72.6%
Moderate	25.6%
Severe	1.8%
Numerical	
Age, yr	9 (8–12)
Height, cm	135 (125–147)
Weight, kg	32 (26–44)
Age of onset of asthma, yr	3 (5–8)
Total IgE	144 (54–375)
Eosinophil, %	3.7 (2.1–6.1)
FEV_1, % predicted	89 (79–99)
FVC, % predicted	101 (92–110)
FEV_1/FVC ratio	85 (80–19)
FEF_{25-75}, % predicted	84 (64–107)
Reversible airway obstruction (n = 243 positive with ⩾12%)	13 (4–16)
Methacholine challenge (n = 95 positive with ⩽8 mg/ml)	1.0 (0.7–2.0)
% Fall in FEV_1 after exercising (n = 83 positive with ⩾10%)	15 (12–20)

Definition of abbreviations: BMI = body mass index; FEF = forced expiratory flow; LABA = long-acting β_2-agonists; SABA = short-acting β_2-agonists.

Categorical variables are given as percentages, and numerical variables are given as median (interquartile range).

multidimensional scaling (29) was then applied to identify relations among different HC methods and deviations from randomization.

The predictive ability of clusters and of original variables with respect to asthma severity (dichotomized into mild vs. moderate–severe) was assessed through information gain ratio (which measures how much information is gained when a variable is known to approximate an outcome) (30), multivariable logistic regression, and Bayesian network analysis (31). Feature selection and network topology optimization were done by stepwise algorithms for both the logistic regression and Bayesian network analyses (32). Specifically, we fitted four logistic models with raw variables: Model 1 included all variables apart from those used to define severity (medication usage, symptoms within the past 4 wk, and FEV_1); Model 2 included all variables; Model 3 was a stepwise selection of variables, adding or removing covariates heuristically based on the Akaike Information Criterion, from Model 1; and Model 4 was a stepwise selection of variables from Model 2. We then reran the logistic regressions, but using cluster memberships as variables. Model performance was assessed by repeated cross-validation and area under the receiver operating characteristic curve (24), which is a composite indicator of sensitivity and specificity. All analyses were performed within the R (www.r-project.org/) and Weka (www.cs.waikato.ac.nz/ml/weka/) software.

The online supplement provides additional details.

RESULTS

Study Population

The characteristics of the study population are shown in Table 1. The cross-sectional set comprised 383 children with asthma, median (interquartile range) age 9 (8–12) years, age of asthma onset of 3 (5–8) years, 60.6% boys, 25.3% classified as obese or overweight, 43.3% with physician-diagnosed allergic rhinitis, 36.0% exposed to tobacco smoke, all receiving SABA, 46.7% receiving additional asthma medication, 17.0% experiencing symptoms within the past 4 weeks, with total serum IgE of 144 (54–375) and FEV_1% predicted of 89% (79–99). Asthma was classified as mild, moderate, or severe in 72.6%, 25.6%, and 1.8% of cases, respectively.

TABLE 2. EIGENVALUES OF THE FIRST 10 COMPONENTS OF THE MULTIPLE FACTOR ANALYSIS AND CONTRIBUTION OF EACH VARIABLE IN THE DATASET TO EACH OF THESE COMPONENTS

		Correlation of Characteristics to Each Dimension				
Group	Variable	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
Lung function	% Predicted FEV_1	−0.41	−0.22	−0.10	−0.12	0.51
	% Predicted FVC	−0.39	0.16	0.10	−0.02	0.39
	FEV_1/FVC ratio	−0.10	−0.48	−0.24	−0.08	0.11
	% Predicted FEF_{25-75}	−0.24	−0.37	−0.18	−0.11	0.35
	Post-bronchodilator FEV_1	−0.17	−0.18	−0.08	0.11	0.16
	Reversible airway obstruction	0.37	0.13	0.12	0.30	−0.34
	Postexercise change in FEV_1	0.09	0.01	−0.10	−0.03	0.32
	PD_{20} methacholine	−0.48	−0.17	−0.08	−0.29	0.13
Exacerbation	Symptoms within last 4 wk	0.40	−0.15	−0.03	0.20	0.40
	Exacerbations within last year	0.39	−0.29	−0.06	0.05	0.49
	Hospitalization	0.21	−0.26	−0.04	0.00	0.39
Family history	Maternal rhinitis	−0.17	−0.05	−0.22	0.37	0.06
	Paternal rhinitis	−0.06	−0.04	−0.56	0.27	−0.04
	Maternal atopy	0.26	0.39	−0.05	−0.25	0.02
	Paternal atopy	0.16	0.31	−0.10	−0.46	0.07
Comorbidities	Rhinitis	−0.12	0.33	−0.38	0.33	0.20
	Eczema	0.08	−0.03	−0.16	0.08	−0.15
	IgE	0.11	0.42	−0.16	0.33	0.09
	% Eosinophils	0.14	0.40	−0.13	0.36	0.18
	Atopy	0.06	−0.52	0.26	−0.30	−0.14
Environment	Exposure to tobacco smoke	0.00	−0.07	0.54	0.52	−0.12
	Pet ownership	−0.23	−0.15	0.48	0.15	0.23
Drugs	ICS	0.60	0.13	0.23	−0.02	0.16
	LABA	0.25	0.41	0.12	−0.05	0.14
	Montelukast	0.27	0.06	0.19	−0.03	0.22
	Any drug in addition to SABA	0.62	0.13	0.23	−0.02	0.19
Growth characteristics	Age of onset of asthma	−0.45	0.20	0.15	0.11	0.13
	Age	−0.41	0.58	0.19	−0.06	0.25
	Sex	0.06	0.29	0.11	0.03	−0.08
	BMI	−0.39	0.29	0.19	−0.16	0.28

Component	Eigenvalue	% of Variance	Cumulative % of Variance
1	1.47	8.50	8.50
2	1.37	7.96	16.46
3	1.13	6.55	23.01
4	1.04	6.03	29.04
5	0.97	5.64	34.68
6	0.91	5.25	39.93
7	0.87	5.04	44.97
8	0.81	4.71	49.68
9	0.81	4.67	54.35
10	0.79	4.55	58.90

Definition of abbreviations: BMI = body mass index; FEF = forced expiratory flow; ICS = inhaled corticosteroids; LABA = long-acting β_2-agonists; SABA = short-acting β_2-agonists.

Exploratory Factor Analysis

The optimal solution from the multiple factor analysis presented five dominant dimensions (eigenvalues ⩾ 1) accounting for only 35% of the total variance of the data. Table 2 shows the eigenvalues of the first 10 dimensions and the correlation of each variable in the dataset to each of these dimensions. The correlation map of variables (Figure 1a) graphically illustrates the correlation of each individual variable with the first principal plane. The significant absolute correlations greater than 0.4 with given dimensions were as follows:

Dimension 1: Medication use (any drug apart from SABA, ICS), lung function (methacholine challenge and FEV_1), age, and age of asthma onset

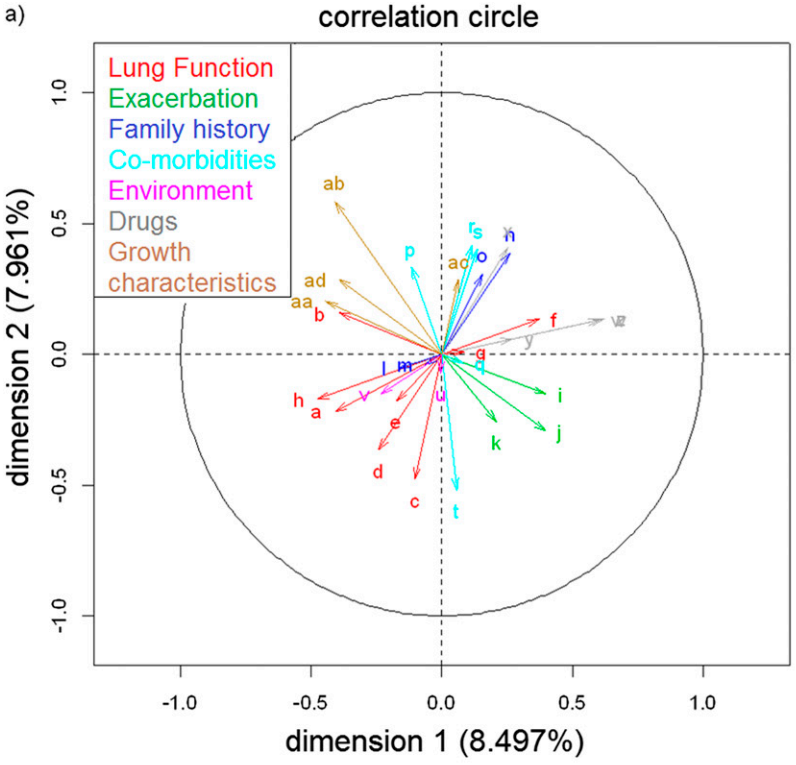
Dimension 2: Age, markers of atopy (IgE, % eosinophils, sensitization), use of LABA, and FEV_1/FVC ratio

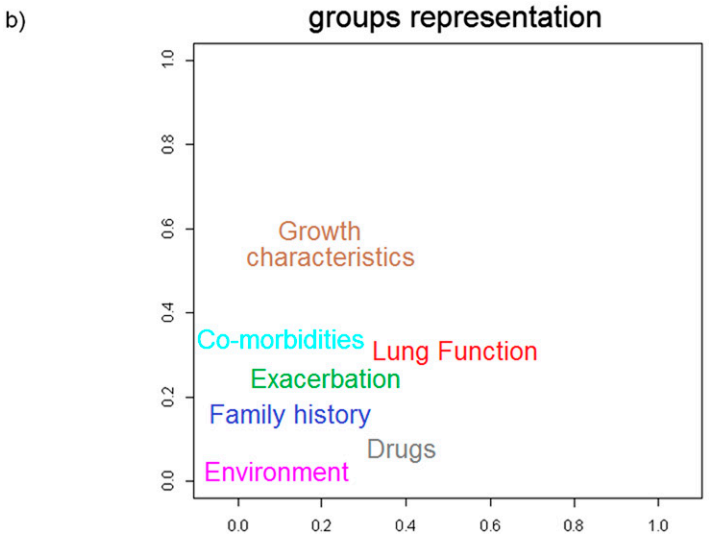
Dimension 3: Rhinitis and environmental exposures (tobacco smoke, pets)

Dimension 4: Paternal atopy and tobacco smoke exposure

Dimension 5: FEV_1 and asthma exacerbations within the past year

Figure 1b shows the coordinates of the imposed groups on the first and second dimensions (16% of variance explained). The plot illustrates that measures of lung function showed a correlation with both Dimensions 1 and 2, use of medication in addition to SABA and exacerbations with Dimension 1, and





Group	Variable	Label
Lung Function	% Predicted FEV_1	a
	% Predicted FVC	b
	FEV_1/FVC Ratio	c
	% Predicted FEF_{25-75}	d
	Post-Bronchodilator FEV_1	e
	Reversible Airway Obstruction	f
	Post-exercise change in FEV_1	g
	PD_{20} Methacoline	h
Exacerbation	Symptoms within last 4 weeks	i
	Exacerbations within last year	j
	Hospitalisation	k
Family history	Maternal Rhinitis	l
	Paternal Rhinitis	m
	Maternal Atopy	n
	Paternal Atopy	o
Co-morbidities	Rhinitis	p
	Eczema	q
	IgE	r
	% Eosinophils	s
	Atopy	t
Environment	Exposure to tobacco smoke	u
	Pet ownership	v
Drugs	ICS	w
	LABA	x
	Montelukast	y
	Any drug in addition to SABA	z
Growth characteristics	Age of onset of asthma	aa
	Age	ab
	Gender	ac
	BMI	ad

Figure 1. (*a*) Principal coordinate plot from the multiple factor analysis. (*b*) Representation of the groups on the first and second dimensions. BMI = body mass index; FEF = forced expiratory flow; ICS = inhaled corticosteroids; LABA = long-acting β_2-agonists; SABA = short-acting β_2-agonists.

general characteristics with Dimension 2. The principal component analysis gave a similar grouping of variables, but with some notable differences in the number of components and the coefficient sets (*see* online supplement). In summary, EFA yielded a low percentage of the variance explained and relatively weak components' characteristics.

Hierarchical Clustering

We performed multiple inferences of HC trees by varying (*1*) the encoding (e.g., binary vs. raw variables), (*2*) the distance-linkage function (e.g., Gower vs. Jaccard distance), and (*3*) the feature selection and dimensionality reduction space. This resulted in a total of 85 trees; we then generated an additional set of 42 random trees (i.e., ratio 2:1).

After identifying clusters, similarity matrices were calculated across all the trees and clusters. There was a clear difference between the trees inferred using the data compared with random trees. On average, real trees produced a lower number of clusters compared with random trees ($P = 0.005$). Real trees were more similar to each other than random trees ($P < 0.0001$); HC plots in Figure 2 (*left*) illustrate segregation between the real trees inferred from the data and the random trees, demonstrating that there is a clear signal in the data.

The variations of the HC method (in the linkage-distance functions) did not affect the cluster inference and yielded similar trees and clusters (Figure 2). However, using different variable encodings and transformations led to more pronounced differences in the clusters, with segregation among real trees.

Prognostic Factors of Severity

After univariate analysis of the raw variables, we ranked them by the information gain ratio in relation to asthma severity (dichotomized into mild vs. moderate–severe) (Figure 3). The highest gain was that of lung function markers and the use of asthma medications in addition to SABA, followed by family history. We next performed multivariable analysis with raw variables (Table 3). Model 1 (in which we excluded FEV_1, current asthma symptoms, and the step of the medication regime, which were used to categorize asthma severity), identified younger age, BMI, paternal asthma, and decreasing FVC and FEV_1/FVC ratio as variables consistently with higher log-odds of moderate–severe asthma. Model 2 (which included all variables) showed that use of asthma medication in addition to SABA (ICS, LABA, or montelukast), asthma symptoms within the last 4 weeks (trend, $P = 0.06$), and lower FEV_1 were significantly associated with moderate–severe asthma, with BMI, and lower FVC and FEV_1/FVC ratio still yielding significant associations. Stepwise Models 3 and 4 selected younger age of asthma onset, lower FVC, FEV_1/FVC ratio, and $FEF_{25–75}$, as associates of moderate–severe asthma (besides medications, symptoms within the past 4 wk, and FEV_1, which were used to define asthma severity). Logistic models using cluster memberships (obtained by HC) as covariates showed consistently worse goodness-of-fit than Models 1–4, both in terms of Akaike Information Criterion and cross-validation estimates using areas under receiver operating characteristic curves (*see* the online supplement).

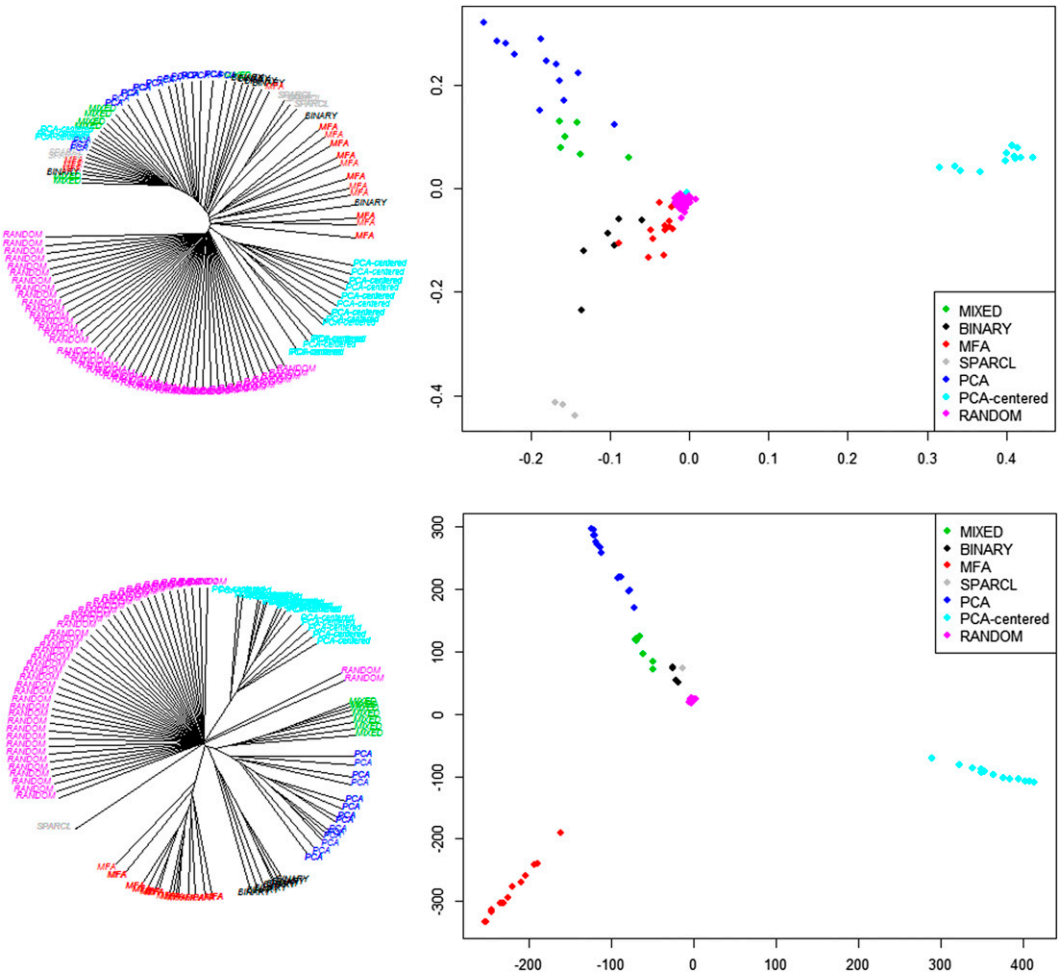


Figure 2. Meta-hierarchical clustering (*left*) and classical multidimensional scaling (*right*) of different clustering methods, compared with randomized trees and clusters, using the adjusted Rand Index (*upper*) and Penny-Hendy tree distance (*lower*). Colors identify different variable encodings, and label replications represent different linkage and distances on the same encoding. The figure shows how there is a clear signal in the data (random trees are segregated from the data clusters), and that variations of the hierarchical clustering method (same color) yield similar trees and clusters (i.e., branching together in a subtree). However, different variable sets, encodings, and transformations lead to more pronounced differences in the clusters.

On deeper analysis using Bayesian network methods we saw a complex conditional structure of variables, which may indicate nonlinear relationships not captured by the logistic models. Figure 4 shows an optimal Bayesian network with these data, with casual dependencies inferred by stepwise algorithm.

DISCUSSION

Summary of Findings and Novelty of Approach

In a cross-sectional study of children with asthma we applied several dimension reduction and data clustering algorithms, associating components, clusters, and raw variables to asthma severity. We systematically explored the effects of varying the variable encodings (e.g., comparing continuous with discretized, or raw variables with those transformed using EFA) and the clustering methods (within HC we tested 85 different models). Changes in linkage and distance resulted in minor changes in the clusters, whereas the changes in the variable encodings and transformations made larger differences to the cluster assignments. Compared with the original raw variables, all of the inferred clusters correlated relatively poorly with asthma severity.

Comparison with Previous Works and Interpretation

Several groups have previously applied different methods of clustering and dimensionality reduction on well-characterized populations of adults and children with asthma to identify patterns within the data. In adults, Moore and coworkers (6) identified five asthma clusters using HC (Ward linkage) of the data from 726 patients with severe asthma. Starting with greater than 600 variables, the data were reduced manually to 34 indicators covering a broad spectrum of routine assessments of asthma without missing data. A subsequent decision tree analysis showed that prebronchodilator and post-bronchodilator FEV_1 and age of onset of asthma were responsible for more than 80% of correct cluster assignments. Haldar and coworkers (7) used the same Ward HC plus a k-means partitioning clustering to infer and compare clusters in two distinct populations (mild–moderate and refractory asthma), validating the findings in a third population of refractory subjects with asthma. However, the variable choice and encoding were different: principal components analysis was performed on 16 variables, chosen among those "considered important in defining the disease phenotype rather than being a product of the disease process" (7) (for instance, post-bronchodilator FEV_1 was not included). Differences among the cluster sets (three in the mild–moderate vs. four in the refractory asthma populations, two in common) for the two populations were discussed in relation to treatment strategies, driven by the role of inflammatory markers. Siroux and coworkers (8) applied latent class analysis on two adult cohorts (n = 641, n = 1,895; 14 markers preprocessed by EFA, including demographics, lung function, and treatment types), showing a discriminatory value of treatment types, a certain degree of stability of inferred clusters (four in both populations, two of these common between the populations), and some resemblance to previous findings. Fitzpatrick and coworkers (9) applied Ward HC to a population of 161 children (>500 variables reduced to 12 by expert advice), and identified four clusters that were highly discriminated by lung function markers, asthma duration, and the use of medications (these parameters were responsible for 93% of the correct cluster assignments). There was some, but not complete resemblance with previous clustering in adults (6), and the clusters had poor discriminating value with respect to asthma severity. Just and coworkers (10) recently identified two novel asthma phenotypes in children (using 19 variables selected by principal components analysis from the initial set of 40) in a three-groups clustering inferred by Ward HC plus a k-means partitioning clustering, focusing on the role of inflammatory markers.

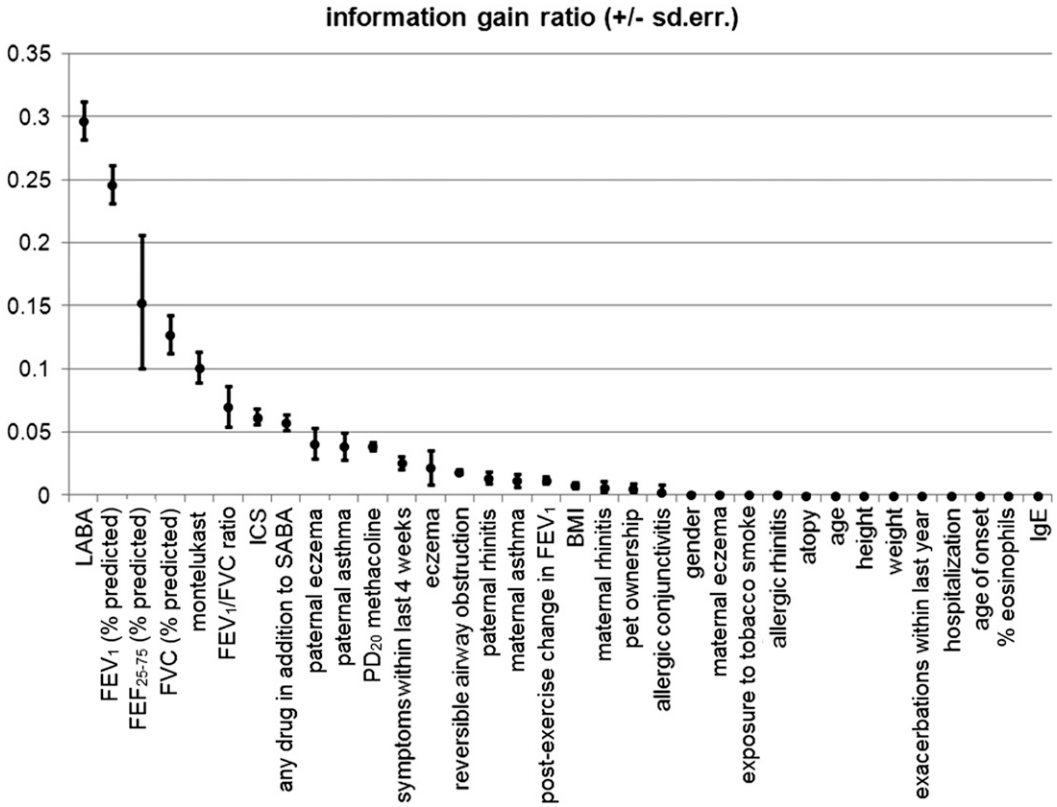


Figure 3. Univariate analysis. Information gain ratio (which measures the % of information gained to approximate the outcome when the variable was known) between single variables and disease severity (mild vs. moderate–severe asthma). BMI = body mass index; FEF = forced expiratory flow; ICS = inhaled corticosteroids; LABA = long-acting β_2-agonists; SABA = short-acting β_2-agonists.

TABLE 3. MULTIVARIABLE ANALYSIS OF THE RAW VARIABLE SET IN RELATION TO ASTHMA SEVERITY (MILD VS. MODERATE–SEVERE)

	Full Covariate Set Models					
	Model 1 (*AIC = 383*)			Model 2 (*AIC = 300*)		
Variable	OR	95% CI	*P* Value	OR	95% CI	*P* Value
Age, yr	1.09	0.97–1.23	0.1336	1.01	0.86–1.19	0.8864
Sex, female vs. male	1.17	0.64–2.16	0.6085	1.15	0.52–2.52	0.7358
Age of asthma onset, yr	0.86	0.78–0.95	0.0031	0.9	0.79–1.03	0.1293
BMI, obese vs. normal	2.11	0.66–6.78	0.2101	1.55	0.38–6.28	0.5389
BMI, overweight vs. normal	0.35	0.14–0.85	0.0208	0.25	0.08–0.8	0.0192
Allergic rhinitis	1	0.5–2	0.9987	0.75	0.3–1.85	0.5273
Allergic conjunctivitis	0.59	0.08–4.32	0.6052	0.5	0.05–4.97	0.5557
Eczema	1.63	0.36–7.43	0.5297	2.44	0.39–15.19	0.3405
Maternal allergic rhinitis	0.72	0.23–2.29	0.5838	1	0.26–3.88	0.9964
Maternal eczema	0.93	0.12–7.03	0.9456	1.3	0.11–15.66	0.8378
Paternal allergic rhinitis	0.48	0.11–2.17	0.3403	0.81	0.11–5.78	0.833
Paternal eczema	3.58	0.43–29.76	0.2375	1.65	0.05–52.19	0.7753
Maternal asthma	0.69	0.17–2.74	0.5982	1.12	0.16–7.73	0.9077
Paternal asthma	3.51	1.16–10.64	0.0264	2.92	0.69–12.39	0.147
Exposure to tobacco smoke	1.34	0.72–2.5	0.355	1.09	0.49–2.39	0.8383
Presence of animals or pets in the home	0.79	0.24–2.59	0.7026	0.8	0.17–3.64	0.769
IgE, per log higher	0.84	0.64–1.09	0.191	0.77	0.55–1.09	0.1387
Eosinophil, %	1.02	0.93–1.12	0.7172	1.03	0.91–1.16	0.6615
Atopy	1.44	0.69–3.03	0.3349	1.19	0.46–3.07	0.7165
FEV_1, % predicted				0.94	0.9–0.98	0.0035
FVC, % predicted	0.93	0.9–0.96	<0.0001	0.95	0.92–0.98	0.0021
FEV_1/FVC ratio	0.9	0.84–0.96	0.0022	0.92	0.85–1	0.0374
FEF_{25-75}, % predicted	0.99	0.97–1.01	0.2101	0.98	0.96–1.01	0.1558
% Fall in FEV_1 after exercise (Q1 vs. Q4)	0.42	0.03–6.27	0.5324	0.09	0–5.52	0.2554
% Fall in FEV_1 after exercise (Q2 vs. Q4)	0.35	0.06–2.21	0.2644	0.53	0.05–5.73	0.6037
% Fall in FEV_1 after exercise (Q3 vs. Q4)	0.51	0.13–2.04	0.3421	0.57	0.09–3.56	0.5515
Bronchodilator reversibility (Q1 vs. Q4)	1.65	0.43–6.4	0.467	3.36	0.58–19.26	0.1745
Bronchodilator reversibility (Q2 vs. Q4)	2.12	0.91–4.97	0.0831	2.18	0.69–6.88	0.1852
Bronchodilator reversibility (Q3 vs. Q4)	0.68	0.3–1.54	0.3529	0.99	0.36–2.71	0.9829
PC_{20} methacholine (Q1 vs. Q4)	2.52	0.19–32.96	0.4803	0.4	0.01–11.9	0.599
PC_{20} methacholine (Q2 vs. Q4)	1.45	0.05–39.25	0.8264	0.08	0–1428.31	0.6191
PC_{20} methacholine (Q3 vs. Q4)	1.49	0.11–19.81	0.7638	1.49	0.07–29.86	0.7961
Asthma symptoms within the last 4 wk				2.52	0.94–6.74	0.0655
No. of exacerbations within the last year	1.05	0.95–1.16	0.3488	1	0.87–1.15	0.9753
No. of hospitalizations	0.94	0.46–1.93	0.868	0.95	0.42–2.16	0.9025
ICS > 300 μg* vs. none				3.58	1.37–9.36	0.0093
ICS ⩽ 300 μg* vs. none				0.98	0.38–2.58	0.9743
LABA				56.33	11.38–278.94	<0.0001
Montelukast				7.08	1.59–31.44	0.0101

	Stepwise Models					
	Model 3 (*AIC = 357*)			Model 4 (*AIC = 260*)		
Stepwise-selected Variable	OR	95% CI	*P* Value	OR	95% CI	*P* Value
Age of asthma onset, yr	0.89	0.81–0.97	0.0057			
FEV_1, % predicted				0.95	0.92–0.97	0.0003
FVC, % predicted	0.93	0.91–0.95	<0.0001	0.96	0.94–0.99	0.0018
FEV_1/FVC ratio	0.88	0.85–0.91	<0.0001			
FEF_{25-75}, % predicted				0.97	0.95–0.98	<0.0001
Asthma symptoms within the last 4 wk				3.01	1.35–6.72	0.0071
ICS >300 μg* vs. none				4.26	1.91–9.5	0.0004
ICS ⩽300 μg* vs. none				1.35	0.6–3.05	0.4694
LABA				34.02	9.51–121.7	<0.0001
Montelukast				6.6	1.69–25.84	0.0067

Definition of abbreviations: AIC = Akaike Information Criterion; BMI = body mass index; CI = confidence interval; FEF = forced expiratory flow; ICS = inhaled corticosteroids; LABA = long-acting β_2-agonists; OR = odds ratio; Q1 = 1st quartile; Q2 = 2nd quartile; Q3 = 3rd quartile; Q4 = 4th quartile; SABA = short-acting β_2-agonists.

Logistic regression models: Model 1 excluded FEV_1, symptoms within the last 4 weeks, and medication step as they were used to define severity; Model 2 included all variables; Model 3 and 4 were stepwise of 1 and 2.

* Betamethasone-equivalent dosage.

In these previous studies of patients with asthma and in similar studies of atopy (33, 34), HC seemed robust with respect to different linkage and distances, or by data bootstrapping. To some extent, it was stable when considering data from different populations with similar variable sets and variable processing. However, our work highlights that the same does not hold when

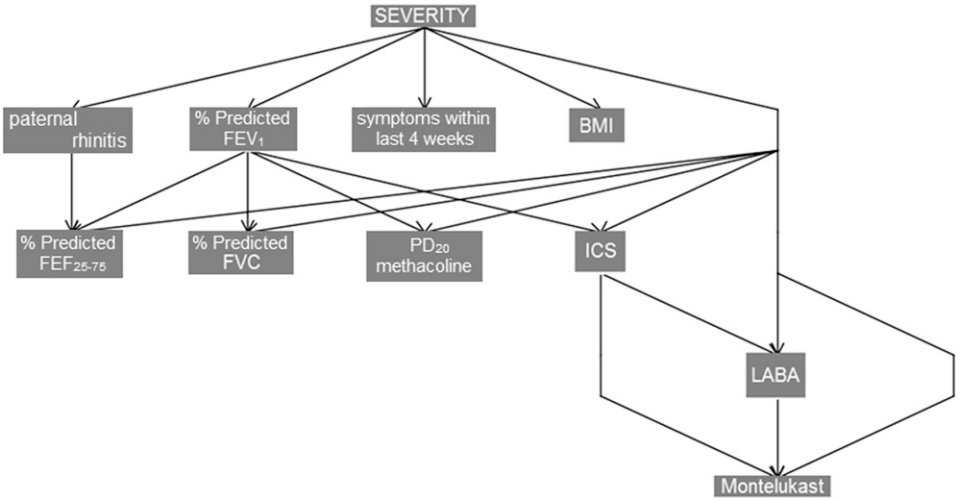


Figure 4. Bayesian network model explaining dependencies between severity as diagnosed by the physician and the original variable space. Both network topology and variables have been selected by a stepwise search. BMI = body mass index; FEF = forced expiratory flow; ICS = inhaled corticosteroids; LABA = long-acting β_2-agonists.

systematically changing the variable sets and encodings, or when transforming the data (e.g., reducing dimensions by EFA), where topologies of trees and cluster sets differ substantially.

The results of our EFA show that there is a diversification of age, age at asthma onset, parental history of disease, environmental factors, lung function markers, exacerbation markers, and inflammatory markers. This diversification was highlighted in previous studies, both in adults (6–8) and children (9, 10). However, the stability of the components was weak, and less robust to changes in the model assumptions (like rotations) or variable discretization policies. The subsequent HC analysis also led to instability in inferred trees and clusters when changing the variable sets, encodings, and transformations.

These findings might be a characteristic exclusive of our study population. An aggregation bias caused by the discretization of skewed variables may play a role, and the use of mixed data types. One previous study (6) highlighted the importance of selecting variables with no missing values, of using normalized variables, and of selecting variable sets explaining the highest variance. Therefore, when asthma subgroups are identified through unsupervised learning, they must be subject to a careful interpretation of the original variable space and its transformations. We did not perform a discrimination analysis of the original variables with respect to each clustering, but this might help to select subsets of variables that lead to more stable clustering, even when varying their encoding.

Consistent with previous findings (9, 10), our HC yielded groups that were relatively poor predictors of asthma severity. This does not imply necessarily that clustering has a poor diagnostic value in general. Indeed, this finding suggests an important point that severe asthma as a phenotype of disease may not be directly associated with unique or uniform pathophysiologic mechanisms (i.e., that it is not a distinct asthma endotype), but likely a phenotypic characteristic at a severe end of the spectrum of a number of asthma subgroups.

When looking at the original variables, prognostic factors of moderate–severe asthma, besides the medication usage, current asthma symptoms, and FEV_1, were paternal asthma, BMI, younger age of asthma onset, and other lung function parameters (FVC, $FEF_{25–75}$, FEV_1/FVC ratio). There was evidence of conditional dependence among variables from the Bayesian network analysis (Figure 4); however, given the high computational complexity of the model selection, the reliability of the network (in terms of variables and relations) was limited by the heuristic procedure for variable selection, and could not be properly quantified.

Methodologic Discussion

In ideal situations, for example with dimension-dense samples of data from normal distributions, different unsupervised learning methods may produce the same results. For instance, it has been demonstrated that the relaxed solution of the k-means clustering algorithm, specified by the cluster indices, is given by the principal components of the data (35). However, even with the same method, different results can be obtained if the analysis is performed with different starting values, or different optimization routines (e.g., using singular value rather than eigenvector decomposition, or using different starting points in k-means). The empirical robustness of a method can be assessed using multiple runs and/or bootstrapping; theoretic robustness, however, may remain debatable. A different case is the conceptual variation of a technique: for instance, in principal component analysis, the promax rotation relaxes the orthogonality assumption, whereas the varimax does not (principal components are guaranteed to be independent only if the dataset is jointly normally distributed).

Medicine often throws up high-dimensional, sparse, noisy data. In such situations, EFA may be applied before HC (36); however, this has been criticized for not being justified in the general case (37). Other approaches include model-based clustering (38). Witten and Tibshirani developed the *sparcl* technique for selecting features in clustering (25), which we used in this study. However, these enhanced methods are difficult to apply in medical research where the typically heterogeneous data (nonnormality, missing values, and mixture of numeric and categorical types) makes it more difficult to design distance metrics (39) or likelihood functions. In addition, different approaches may have different ways of identifying the optimal number of clusters by their internal measures (40) or external indices (41), in which case ensemble approaches (42) might be needed.

Conclusions

Unsupervised statistical learning can help investigators to identify complex patterns and structures in data, and to reduce dimensionality to something conceivable. This interaction with

the data may in turn generate or shape novel hypotheses. However, the patterns used for hypothesis generation must be reliable. We have shown that clustering using different variable sets and encodings in asthma datasets can lead to different clusters. A more thoughtful selection of markers, encoded appropriately, and consistent across different populations is required before attempting unsupervised statistical learning. Then, careful interpretation of the variable space and its transformations are essential if true asthma subgroups are to be identified by interacting with data in this way.

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1. Papierniak ES, Lowenthal DT, Harman E. Novel therapies in asthma: leukotriene antagonists, biologic agents, and beyond. *Am J Ther* 2013; 20:79–103.
2. Bacharier LB, Guilbert TW. Diagnosis and management of early asthma in preschool-aged children. *J Allergy Clin Immunol* 2012;130:287–296; quiz 297–288.
3. Lötvall J, Akdis CA, Bacharier LB, Bjermer L, Casale TB, Custovic A, Lemanske RF Jr, Wardlaw AJ, Wenzel SE, Greenberger PA. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol* 2011;127:355–360.
4. Sittka A, Vera J, Lai X, Schmeck BT. Asthma phenotyping, therapy, and prevention: what can we learn from systems biology? *Pediatr Res* 2013;73:543–552.
5. Custovic A, Lazic N, Simpson A. Pediatric asthma and development of atopy. *Curr Opin Allergy Clin Immunol* 2013;13:173–180.
6. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R Jr, Castro M, Curran-Everett D, Fitzpatrick AM, *et al.*; National Heart, Lung, and Blood Institute's Severe Asthma Research Program. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 2010;181:315–323.
7. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, Wardlaw AJ, Green RH. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008;178:218–224.
8. Siroux V, Basagana X, Boudier A, Pin I, Garcia-Aymerich J, Vesin A, Slama R, Jarvis D, Anto JM, Kauffmann F, *et al.* Identifying adult asthma phenotypes using a clustering approach. *Eur Respir J* 2011;38: 310–317.
9. Fitzpatrick AM, Teague WG, Meyers DA, Peters SP, Li X, Li H, Wenzel SE, Aujla S, Castro M, Bacharier LB, *et al.* Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *J Allergy Clin Immunol* 2011;127:382–389, e1–e13.
10. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. *Eur Respir J* 2012; 40:55–60.
11. Smith JA, Drake R, Simpson A, Woodcock A, Pickles A, Custovic A. Dimensions of respiratory symptoms in preschool children: population-based birth cohort study. *Am J Respir Crit Care Med* 2008;177:1358–1363.
12. 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, e12.
13. Saraçlar Y, Sekerel BE, Kalayci O, Cetinkaya F, Adalioğlu G, Tuncer A, Tezcan S. Prevalence of asthma symptoms in school children in Ankara, Turkey. *Respir Med* 1998;92:203–207.
14. Stanojevic S, Wade A, Cole TJ, Lum S, Custovic A, Silverman M, Hall GL, Welsh L, Kirkby J, Nystad W, *et al.*; Asthma UK Spirometry Collaborative Group. Spirometry centile charts for young caucasian children: the Asthma UK Collaborative Initiative. *Am J Respir Crit Care Med* 2009;180:547–552.
15. Crapo RO, Casaburi R, Coates AL, Enright PL, Hankinson JL, Irvin CG, MacIntyre NR, McKay RT, Wanger JS, Anderson SD, *et al.* Guidelines for methacholine and exercise challenge testing-1999. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. *Am J Respir Crit Care Med* 2000;161:309–329.
16. Sekerel BE, Saraçlar Y, Kalayci O, Cetinkaya F, Tuncer A, Adalioğlu G. Comparison of four different measures of bronchial responsiveness in asthmatic children. *Allergy* 1997;52:1106–1109.
17. Tahan F, Karaaslan C, Aslan A, Kiper N, Kalayci O. The role of chemokines in exercise-induced bronchoconstriction in asthma. *Ann Allergy Asthma Immunol* 2006;96:819–825.
18. Vidmar S, Carlin J, Hesketh K, Cole T. Standardizing anthropometric measures in children and adolescents with new functions for egen. *Stata J* 2004;4:50–55.
19. Cropp GJ. The exercise bronchoprovocation test: standardization of procedures and evaluation of response. *J Allergy Clin Immunol* 1979;64:627–633.
20. Joos GF, O'Connor B, Anderson SD, Chung F, Cockcroft DW, Dahlen B, DiMaria G, Foresi A, Hargreave FE, Holgate ST, *et al.* Indirect airway challenges. *Eur Respir J* 2003;21:1050–1068.
21. Custovic A, Arifhodzic N, Robinson A, Woodcock A. Exercise testing revisited. The response to exercise in normal and atopic children. *Chest* 1994;105:1127–1132.
22. Becue-Bertaut M, Pages J. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Comput Stat Data Anal* 2008;52:3255–3268.
23. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken, NJ: Wiley; 2005.
24. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction. New York: Springer; 2009.
25. Witten DM, Tibshirani R. A framework for feature selection in clustering. *J Am Stat Assoc* 2010;105:713–726.
26. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008;24:719–720.
27. Hubert L, Arabie P. Comparing partitions. *J Classif* 1985;2:193–218.
28. Penny D, Hendy MD. The use of tree comparison metrics. *Syst Zool* 1985;34:75–82.
29. Cox TF, Cox MAA. Multidimensional scaling. Boca Raton, FL: Chapman & Hall/CRC; 2001.
30. Quinlan JR. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.
31. Pearl J. Causality: models, reasoning, and inference. Cambridge, UK: Cambridge University Press; 2000.
32. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Mach Learn* 1992;9:309–347.
33. Simpson A, Tan VY, Winn J, Svensén M, Bishop CM, Heckerman DE, Buchan I, Custovic A. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med* 2010;181:1200–1206.
34. Lazic N, Roberts G, Custovic A, Belgrave D, Bishop CM, Winn J, Curtin JA, Hasan Arshad S, Simpson A. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* 2013;68:764–770.
35. Zha HY, He XF, Ding C, Simon H, Gu M. Spectral relaxation for k-means clustering. In: Dietterich TG, Becker S, Ghahramani Z, editors. Advances in neural information processing systems 14: proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference, Vol 14, Part 2. Cambridge, MA: MIT Press; 2002.
36. Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* 2002;18:275–286.
37. Chang W-C. On using principal components before separating a mixture of two multivariate normal distributions. *J R Stat Soc Ser C Appl Stat* 1983;32:267–275.
38. Raftery AE, Dean N. Variable selection for model-based clustering. *J Am Stat Assoc* 2006;101:168–178.
39. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics* 1971;27:857–871.

40. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster-analysis. *J Comput Appl Math* 1987;20: 53–65.
41. Färber I, Günnemann S, Kriegel H-P, Kröger P, Müller E, Schubert E, Seidl T, Zimek A. On using class-labels in evaluation of clusterings. Presented at the 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust 2010) in conjunction with 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010), July 25–28, 2010, Washington, DC.
42. Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 2003;3:583–617.