

Abbreviations used

AQLQ:	Asthma Quality of Life Questionnaire
BAL:	Bronchoalveolar lavage
FENO:	Fraction of exhaled nitric oxide
HC:	Healthy control subject
HCU:	Health care use
INFOGAIN:	Information gain
SA:	Severe asthma
SARP:	Severe Asthma Research Program

with 2 specifically including extremely well-characterized adults with severe asthma (SA).^{2,3} The Leicester study used a *k*-means clustering approach³ that was limited to 16 variables but included sputum eosinophil counts to identify 4 severe and 2 mild-to-moderate asthma clusters. Moore et al² performed a hierarchical clustering of asthmatic patients from the Severe Asthma Research Program (SARP), reducing 628 variables to 34 core variables to identify 5 clusters. Variables related to inflammatory markers were not included. Thus additional approaches are needed incorporating greater numbers of variables, inflammatory markers, or both.

Machine learning techniques have recently been applied to human diseases. Algorithms have assisted in selecting features from thousands of genes to facilitate biomarker identification and accurate patient diagnosis. For example, a feature selection framework based on information theory involves ranking features by using a correlation measurement information gain (INFOGAIN) and then selecting relevant and nonredundant features by using a Markov blanket algorithm.⁹ This framework was applied to a leukemia microarray data set to obtain subsets of nonredundant features from 7130 genes that distinguished acute lymphocytic from acute myeloid leukemia with high accuracy.¹⁰ When the top 3 informative and nonredundant genes were selected to predict leukemia types, 100% classification accuracy was achieved. Therefore machine learning techniques offer promising approaches to understand complex diseases.

Although a clustering analysis was previously performed in SARP, blood/bronchoscopic inflammatory characteristics and IgE and fraction of exhaled nitric oxide (FENO) values were not included because the majority had not undergone these procedures/tests. Additionally, healthy control subjects (HCs) were excluded because the majority of variables analyzed were clinically related to asthma. This second SARP analysis performed an unsupervised clustering of more than 100 variables on 378 asthmatic patients and HCs who had undergone bronchoscopy to incorporate inflammatory variables. A feature selection framework selected 51 relevant and nonredundant variables. Predictive values of the full set of 112 variables and the selected 51 variables were assessed for their ability to distinguish subject clusters.

METHODS**Patient population**

Subjects were all part of SARP. Characterization of the subjects can be found in the [Methods](#) section in this article's Online Repository at www.jacionline.org. SA was defined by 2000 American Thoracic Society workshop criteria (see the [Methods](#) section in this article's Online Repository).¹¹ Other asthmatic patients were divided into 4 groups on the basis of FEV₁ percent predicted values and inhaled corticosteroid use, as previously described.¹²

HCs from SARP were included as well. This clustering was limited to subjects with "lung" variables, including bronchoalveolar lavage (BAL) fluid cell counts and FENO values (see the [Methods](#) section in this article's Online Repository).

Computational and statistical analysis

Data preprocessing. Variables with 5% or greater missing data were excluded. Missing values in variables with less than 5% missing data were added by using a k-NNimpute algorithm.¹³

Cluster analysis. A *k*-means clustering method was applied to the preprocessed data to partition subjects (including HCs). A Ward agglomerative hierarchical clustering method applied to the preprocessed data grouped clinical variables.

Statistical tests. Data for continuous variables were log transformed to improve normality of distribution. Welch *t* statistics,¹⁴ which allow for data in different groups with unequal variances and are popularly used in microarray analysis,¹⁵ were used for log₂-transformed data from continuous variables. χ^2 Tests were used for data from categorical variables to find differences between subjects in the total SARP cohort and those in this cluster analysis. ANOVA and pairwise *t* tests were used for continuous variables with log₂-transformed data, Kruskal-Wallis and pairwise Wilcoxon rank sum tests were used for categorical ordinal variables, and χ^2 tests and pairwise Fisher exact tests were used for categorical binary variables to find significant differences between subject clusters. *P* values were adjusted for multiple testing by using a false discovery rate procedure.¹⁶ *P* values of less than .05 were considered significant. Data presented in the text and figures are on the original scale.

Ranking variables by using INFOGAIN. INFOGAIN¹⁷ measures how well a variable predicts subject cluster labels, as determined by using *k*-means clustering. Higher INFOGAIN values identify which variables best predict subject cluster labels. Calculating INFOGAIN values requires continuous variables to undergo discretization. After calculating their INFOGAIN values, variables were ranked according to their values.

Selection of relevant and nonredundant variables. A set of relevant variables was selected if their INFOGAIN values were greater than a threshold of 0.05; redundant variables were removed by using a Markov blanket algorithm.¹⁸

Classification of clustered subjects. Subjects whose cluster labels were assigned by means of *k*-means clustering were classified by using a multiclass support vector machine algorithm. The 378 subjects were split into a training data set (80% subjects) and a test data set (20% subjects). Leave-one-out cross-validation was performed to estimate classification errors.

Further details for the following methods can be found in the [Methods](#) section in this article's Online Repository: data preprocessing, cluster analysis, statistical tests, discretization and ranking variables by using INFOGAIN, selection of relevant nonredundant variables, and classification of clustered subjects.

RESULTS**Demographics**

Only 378 subjects (of 1685 total) with BAL, blood, and FENO data were analyzed. These subjects, including HCs and patients with mild asthma, moderate asthma, and SA, did not generally differ from the total cohort (see [Table E1](#) in this article's Online Repository at www.jacionline.org) but were older and with slightly better lung function likely caused by limitations on research bronchoscopy in subjects less than 18 years old or with more severe obstruction.

Clustering results

Three hundred seventy-eight subjects were clustered into 6 groups ([Table I](#)). One hundred twelve variables were clustered

TABLE I. Demographic data of subjects in the subject clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	P value* (FDR)
Age at enrollment	29 ± 11†	31 ± 11	33 ± 10	39 ± 12	47 ± 10	43 ± 11	<.0001
Male/female (%)	41/59	43/57	16/84	25/75	60/40	25/75	.002
Caucasian/African American/other (%)	79/9/12	77/19/4	42/11/47	53/41/6	93/7/0	67/33/0	<.0001
Hispanic/non-Hispanic/unknown (%)	4/96/0	1/99/0	63/0/37	0/100/0	3/97/0	0/96/4	<.0001
Body mass index (kg/m ²)	26 ± 7	27 ± 5	29 ± 9	31 ± 7	31 ± 6	30 ± 7	<.0001
Prebronchodilator FEV ₁ (% predicted)	100 ± 11	88 ± 15	93 ± 13	71 ± 14	65 ± 17	50 ± 21	<.0001
Severity level (no.)‡							<.0001
HCS	89						
Patients with mild asthma/no ICS	2	49	2	4			
Patients with mild asthma + ICS		47	6	8	2		
Patients with moderate asthma/no ICS		13		6	1		
Patients with moderate asthma + ICS		17	1	18	1		
Patients with SA		9	10	43	26	24	
Total	91	135	19	79	30	24	

ICS, Inhaled corticosteroid.

*False discovery rate (FDR)—adjusted *P* value from ANOVA or χ^2 tests among 6 subject clusters.†Numeric data are presented as antilog of log₂-transformed means ± SDs.‡See the [Methods](#) section for a detailed description.

into 10 categories by using cluster analysis (see [Fig E1](#) and [Table E2](#) in this article's Online Repository at www.jacionline.org).

Ranking variables

INFOGAIN values were calculated for each variable (see [Table E3](#) in this article's Online Repository at www.jacionline.org), demonstrating that the most discriminating variable was age of asthma onset.^{2,3,19,20} Of the top 10 variables, 9 were symptom and medication related and 1 was asthma history related (age of asthma onset). By using an INFOGAIN value of 0.2 (of a possible 1.0) as a cutoff, 45 variables distinguished the subject clusters: 16 symptom-related, 10 medication-related, 10 asthma history/comorbidities-related, 7 allergy-related, 5 lung function-related, and 3 health care use (HCU)-related variables.

Variable clustering

Statistical tests showed 101 of 112 variables differed among subject clusters (see [Table E3](#)), suggesting these variables represent characteristics distinguishing asthmatic patients from HCs, as well as distinguishing asthma phenotypes. Hierarchical clustering partitioned the 112 variables into 10 clusters, which were then ordered by the median INFOGAIN value for each cluster (see [Table E2](#)). The dendrogram of the variables from cluster analysis is shown (see [Fig E2](#) in this article's Online Repository at www.jacionline.org).

Variable cluster 1. This cluster ([Fig 1](#)) has the highest median INFOGAIN value (0.68), indicating these variables best distinguish the subject clusters. This cluster contains symptom and Asthma Quality of Life Questionnaire (AQLQ) domain variables (see [Table E2](#)), indicating AQLQ scores correlate highly with symptoms. Three AQLQ variables (Symptoms, Emotion, and Environment) are redundant for the activity domain (see [Table E4](#) in this article's Online Repository at www.jacionline.org). These variables distinguish HCs from asthmatic patients, as well as the various asthma clusters, contributing to their high INFOGAIN values (see [Table E3](#)).

Variable cluster 2. Medication and atopy/allergy variables characterize this cluster ([Fig 2](#)). This cluster is anchored by the variable with the strongest predictive value: age of asthma onset

(0.84). Eight variables, including triggers for asthma, asthma medication use, and atopy, only distinguish HC clusters (subject cluster 1) from asthmatic clusters. Despite having modest-to-high INFOGAIN values (most >0.19), many variables are redundant with age of asthma onset (see [Table E4](#)), suggesting they have less impact on subject clusters when the latter is included.

Variable cluster 3. Four lung function variables characterize this cluster (see [Fig E3](#) in this article's Online Repository at www.jacionline.org), with prebronchodilator FEV₁ percent predicted having the highest INFOGAIN value (0.48). Overall, these variables have a moderate INFOGAIN value (0.29), supporting a moderate ability to separate asthmatic patients from HCs and asthma phenotypes.

Variable cluster 4. This cluster ([Figs 3 and 4](#)) contains medication variables, primarily related to corticosteroids, and HCU variables. This cluster is characterized by frequent and severe HCU, as well as nasal polyps, nasal-sinus disease, and history of reactions to aspirin and other nonsteroidal pain medications. It is likely that these clinical characteristics necessitate greater need for systemic corticosteroids, explaining their close relationship in this cluster.²¹ Finally, osteoporosis also falls into this cluster, likely a consequence of high-level corticosteroid use.

Variable cluster 5. This cluster ([Fig 5](#)) contains 3 additional lung function variables and blood and lung T_H2 inflammation-related variables. Even though these variables correlated with lung volume variables in cluster 3 (see [Fig E3](#)), the 3 lung function variables in this cluster correlated more strongly with these inflammatory variables than with cluster 3 lung function variables. FENO has the highest INFOGAIN value (0.16) of the inflammatory variables.

Variable cluster 6. This cluster (see [Fig E4](#) in this article's Online Repository at www.jacionline.org) contains noneosinophilic blood inflammatory markers, including total white blood cell counts and neutrophil numbers/percentages in blood. The overall INFOGAIN value is modest (0.12).

Variable cluster 7. This cluster contains age, number of children, body mass index, and systolic and diastolic blood pressure, all of which could be related to age.

Variable cluster 8. This cluster contains additional lung inflammatory variables. Interestingly, the INFOGAIN value of

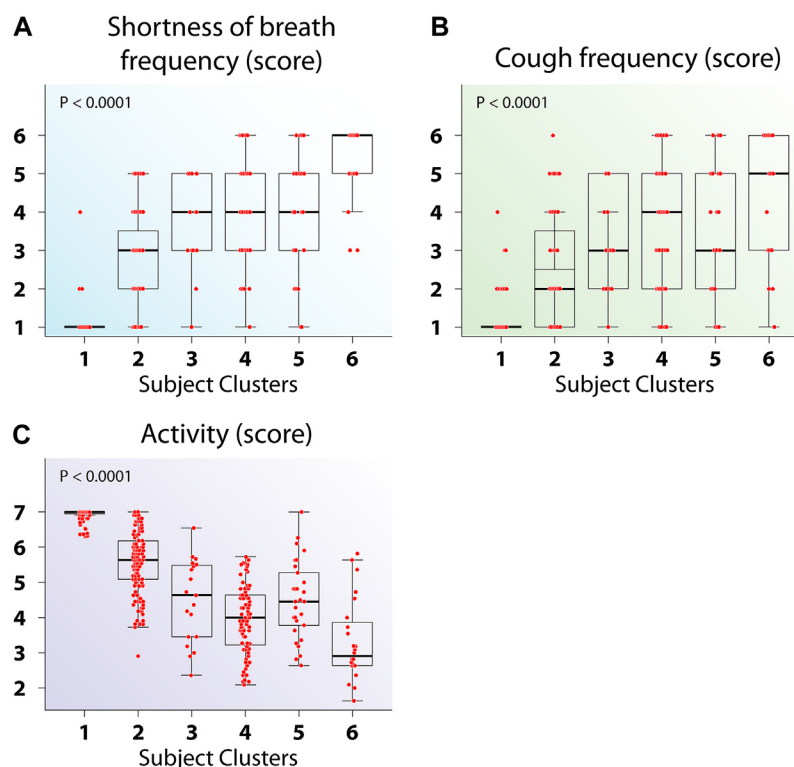


FIG 1. Asthma symptom and quality-of-life variables differed by subject clusters. Cluster 6 had the highest shortness of breath frequency (**A**), the highest cough frequency (**B**), and the lowest AQLQ activity limitation score (**C**) among all clusters. The intergroup differences for all variables can be found in [Table E3](#).

these lung variables (overall 0.07) is less than that of the same variables in blood (0.12).

Variable clusters 9 and 10. These clusters have the lowest median INFOGAIN values among the clusters. The strongest factors in each cluster are height (cluster 9; INFOGAIN value, 0.11) and non-Hispanic ethnicity (cluster 10; INFOGAIN value, 0.23), respectively. Environmental factors and immune and family history variables appear here, although with small impact.

Subject clustering

Analysis of these 10 clusters of clinical, immune, inflammatory, and physiologic variables produced 6 subject clusters. Intergroup statistical comparisons determined the variables that differed significantly across subject clusters (listed in the Figure legends and [Table E3](#)).

Subject cluster 1 (n = 91). All but 2 subjects were HCs with normal lung function and no symptoms, implying these 2 asthmatic patients were more similar to HCs than the other asthmatic patients. Cluster 1 differed from all other clusters in 38 of 112 variables but did not differ from cluster 2 in 7 additional variables, including intensive care unit admissions, hospitalizations, oral corticosteroids, nebulized β -agonist use, and blood monocyte counts. Instances in which cluster 1 differs from the other clusters beyond these specific differences are noted in the text or figure legends.

Subject cluster 2 (n = 135). Cluster 2, the largest cluster, is primarily patients with mild asthma, although 29% have predefined moderate-to-severe asthma. Despite these different traditional severity levels, cluster 2 subjects exhibit similar patterns across variables, with less asthma symptoms and better

AQLQ scores than the remaining clusters ([Fig 1](#)). They have early-onset disease and fewer winter allergy symptoms but numerous allergen skin test reactions ([Fig 2](#)). They have slightly but significantly lower lung function than the “normal” cluster 1 but better lung function than clusters 4 to 6, especially after bronchodilator ([Fig 5](#) and see [Fig E3](#)). They are unlikely to have frequent HCU, high corticosteroid use, nasal polyps, and corticosteroid-related side effects ([Figs 3 and 4](#)). Cluster 2 has higher blood and BAL eosinophil and neutrophil counts than cluster 1 ([Fig 5](#) and see [Figs E4 and E5](#) in this article’s Online Repository at www.jacionline.org).

Subject cluster 3 (n = 19). Cluster 3 is predominantly Hispanic women, with 53% predefined as having SA ([Table I](#) and see [Fig E6, A](#), in this article’s Online Repository at www.jacionline.org) and frequent symptoms, low AQLQ scores, a high degree of allergic sensitization, and likely to have seen a doctor for asthma in the previous year ([Figs 1-3](#)). Despite this, they have near-normal FEV₁ percent predicted values, which reverse to normal after bronchodilator, and little inflammation. (Blood eosinophil counts and FENO values did not differ from cluster 1; BAL eosinophil counts were higher than in cluster 1 but lower than in cluster 5 [[Fig 5](#) and see [Figs E3 and E5](#)]).

Subject cluster 4 (n = 79). Cluster 4 is also predominantly female (75%), 47% nonCaucasian ([Table I](#)), and 54% have predefined SA. They have a high body mass index ([Table I](#)), frequent asthma symptoms, an early age of asthma onset, and are likely to have seen a doctor for asthma in the previous year ([Figs 1-3](#)). Forty-nine percent report “either mother or father with asthma” (see [Fig E6, B](#)), suggesting these subjects are genetically predisposed to asthma. Their FEV₁ is lower than that seen in clusters 2 and 3 but also reverses to near normal (see [Fig E3](#)).

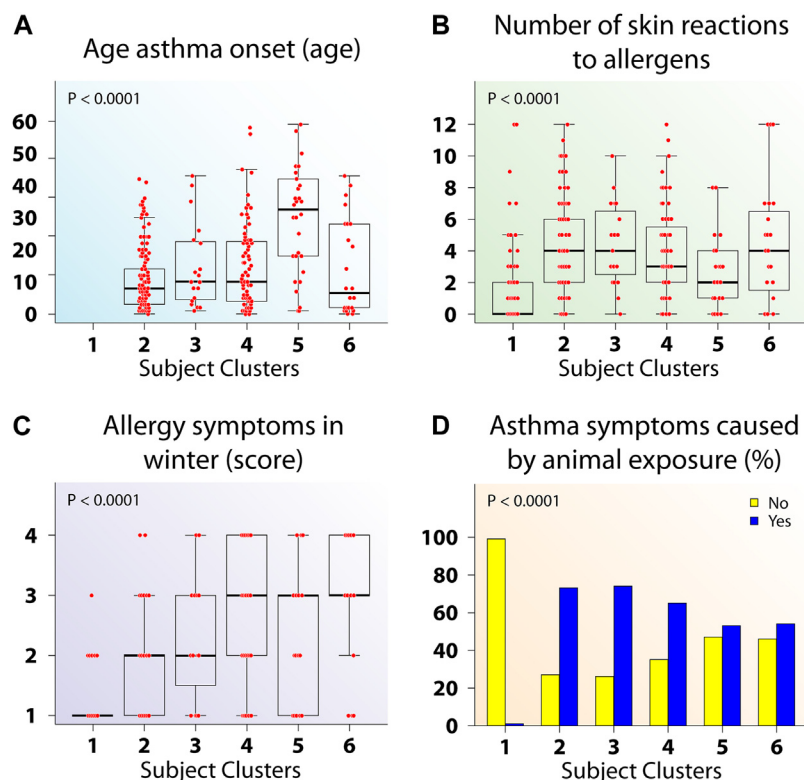


FIG 2. Age of asthma onset and allergy differed by subject clusters according to asthma disease status. **A** and **B**, Cluster 5 had later onset (Fig 2, **A**) and lower numbers of allergen-induced skin reactions (Fig 2, **B**) than all other clusters. **C**, Cluster 2 had less allergy symptoms in winter than clusters 4 to 6. **D**, Clusters 2 to 6 have more asthma symptoms caused by animal exposure than cluster 1.

Cluster 4 subjects have higher FENO values and blood and BAL eosinophil and neutrophil counts than cluster 1 but these do not differ from clusters 2 and 3 (Fig 5 and see Figs E4 and E5).

Subject cluster 5 (n = 30). Cluster 5 contains mostly patients with SA, with 4 classified as having mild or moderate asthma (Table I). This cluster is older (age, 47 years), with later age of asthma onset and lower numbers of allergic skin reactions but a strong history of nasal polyps, sinusitis, and high-dose corticosteroid use (Figs 2 and 4). This cluster has poor lung function, blood and BAL eosinophilia, and increased total white blood cell and blood neutrophil counts (Fig 5 and see Figs E3-E5). Activity limitation is less than that seen in the next cluster (cluster 6, Fig 1).

Subject cluster 6 (n = 24). Cluster 6 consists of only patients with early-onset SA, the most symptoms, the lowest lung function, frequent and high-intensity HCU, and sinusitis (Figs 1-5 and see Fig E3). Despite high oral corticosteroid use (Fig 4), these subjects also have the highest FENO values, high BAL eosinophil counts, more osteoporosis, and higher blood neutrophil counts (Figs 4 and 5 and see Figs E4 and E5). This combination suggests compliance with corticosteroids (blood neutrophilia and osteoporosis) but poor lung-specific anti-inflammatory responses (persistently increased FENO values and BAL eosinophil counts).

Comparison of current subject clusters with previous SARP clusters

The previous SARP cluster analysis used 34 variables without including inflammatory variables to identify 5 asthma phenotypes

among more than 700 total participants. Comparing these 2 studies (see Table E5 in this article's Online Repository at www.jacionline.org), subject cluster 2 in our analysis largely overlapped with the milder clusters (previous SARP clusters 1 and 2). Cluster 3, primarily a mix of previous clusters 1 and 2, identified a group of early-onset atopic Hispanic women with high symptom burden despite normal lung function and little overlap with the predominantly female, late-onset cluster identified in the earlier SARP clusters. Our cluster 4 generally overlapped with the moderate-to-severe asthma clusters 2 to 4, while our subject cluster 6 overlapped with the most severe clusters 4 and 5 in the previous study. Similar to our cluster 3, our cluster 5 showed little overlap with any cluster in the previous study. These results suggest general similarities across analyses, particularly in patients with mild and very severe asthma, but also distinct definition of 2 additional clusters.

Variable predictive values

Subject clusters were classified by using all variables, as well as the INFOGAIN top 10 variables, to investigate the predictive values of these variables (see the Results section in this article's Online Repository at www.jacionline.org). Subjects were split into a training data set containing 80% and a test data set containing 20% of the subjects for multiple times. By using a multiclass support vector machine classifier with the full set of variables, the prediction accuracy rate on the test data was $93\% \pm 2\%$ on average, whereas the prediction accuracy rate decreased to $73\% \pm 5\%$ by using the top 10 variables, supporting

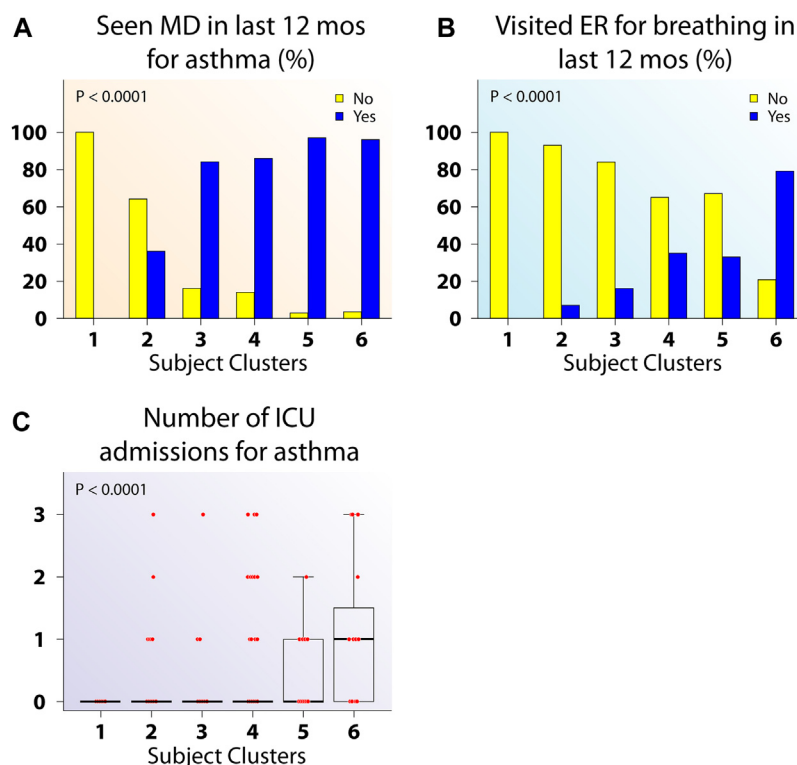


FIG 3. HCU differed by subject clusters. **A**, Subjects in clusters 3 to 6 were more likely to have seen a doctor in the last 12 months for asthma than those in cluster 2. **B** and **C**, Cluster 6 had a higher proportion of subjects who visited the emergency department (ER) for breathing in the last year (Fig 3, **B**) and number of intensive care unit (ICU) admissions for asthma (Fig 3, **C**) than all other clusters.

the importance of variables such as lung function and inflammatory factors in distinguishing different clusters.

The impact of removing redundant variables was also addressed. A Markov blanket algorithm combined with INFOGAIN selected 51 relevant nonredundant variables (see Table E4). Age of asthma onset, quality of life, symptoms, medication use, HCU, nasal polyps, allergy, lung function, and inflammatory factors are the major features distinguishing different clusters. The prediction rate on the test data set was $88\% \pm 3\%$, performing less accurately than when including all variables and suggesting that redundant variables carry extra valuable information.

DISCUSSION

In this study unsupervised machine learning approaches generated 10 clusters of clinical, physiologic, and immunoinflammatory variables, categorizing 6 subject clusters from 378 SARP asthmatic and HC participants with bronchoscopic inflammatory and FENO data. By using the INFOGAIN value as a nonlinear correlation measure and a Markov blanket algorithm, variables were ranked by their ability to separate subject clusters, identifying 51 nonredundant variables. Despite inclusion of inflammatory variables, variables for age at asthma onset, quality of life, symptoms, medication use, and HCU were the top nonredundant features, likely because of their ability to differentiate HCs from asthmatic patients but also their ability to distinguish asthma clusters. Lung function variables were moderately discriminatory, as were allergic symptoms, nasal polyps, and inflammatory profiles, being less distinguishing

overall but important for differentiating 1 or 2 subject clusters well. The inclusion of these variables likely contributed to identification of clusters 3 and 5, which, in contrast to the previous SARP clusters, overlap with similar clusters in the Leicester cohort.³ Thus this clustering approach brings novel insights to asthma phenotyping while helping confirm and integrate phenotypes of previous reports.

One obvious difference between this and previous analyses is inclusion of HCs, which promoted the most statistically unsupervised perspective possible, including “all comers” and not differentiating by clinically biased diagnosis. This analysis included blood, BAL and exhaled breath measures, which are not specific for asthma, overlapping substantially with HCs. Analyzing their distribution among clusters was not “biased” toward differences in asthma groups, allowing assessment of their clustering with clinical variables. By using this approach, 2 subjects classified as “asthmatic” clustered in “HC” cluster 1, having more similarity to HCs than any asthmatic cluster, likely because of their normal lung function and minimal symptoms. Although both subjects had modest methacholine hyperresponsiveness, these data were not included because of substantial missing data, especially in patients with SA. Although excluding HCs yielded almost identical asthmatic cluster results, including HCs allows comparisons of inflammatory patterns (and other variables) that contribute to overall relationships.

All 112 variables were required to achieve nearly perfect test set classification results (93% accuracy compared with 88% accuracy with the 51 nonredundant variables). Despite their small relative INFOGAIN values, redundant variables carry additional

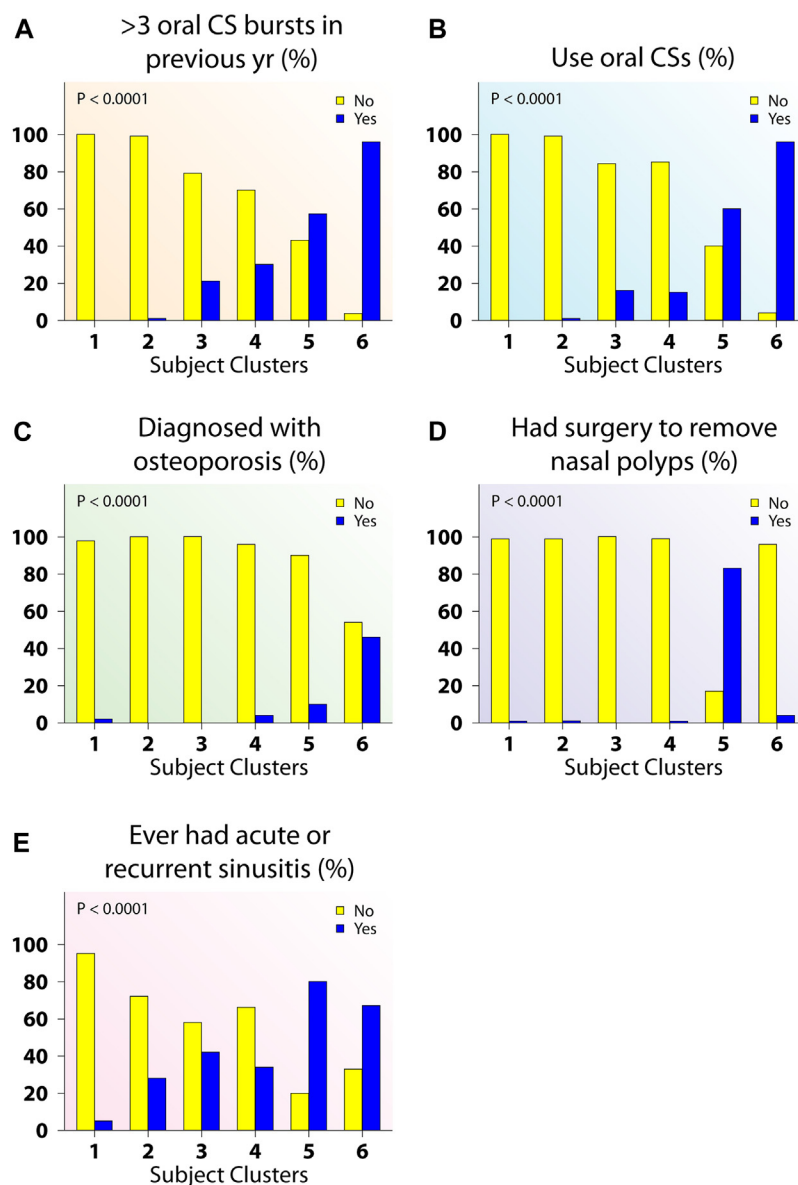


FIG 4. Corticosteroid (CS) use, treatment consequences, and associated clinical characteristics differ across subject clusters. **A-C**, Cluster 6 had the higher proportion of subjects with more than 3 oral corticosteroid bursts in the previous year (Fig 4, *A*), receiving oral corticosteroids (Fig 4, *B*), and having osteoporosis (Fig 3, *C*) than all other asthma clusters. **D** and **E**, Cluster 5 had the highest proportion of subjects with nasal polyps removed (Fig 4, *D*) and with sinusitis (Fig 4, *E*).

important information. Clinically, it would be impossible to use 112 variables to classify subjects. Thus investigations are needed to identify smaller sets of variables that perform as well as the full set and ideally determine whether novel biomarkers might identify these clusters.

This is the third cluster approach analyzing adults with SA and the second cluster approach on the adult SARP population. Overlapping with 250 previous SARP cluster analysis subjects, there was reasonable agreement in the mildest and most severe clusters. The current approach included more variables and immunoinflammatory characteristics. These differences likely contributed to identification of “new” cluster 5 (older onset and mostly SA with nasal polyps), as well as the rather distinct cluster 3. Cluster 5 is recognizable in some form as

aspirin-exacerbated respiratory disease. However, sensitivity to “aspirin or other nonsteroidal anti-inflammatory drugs” appears less critical to this phenotype than the presence of nasal polyps and sinusitis in patients with late-onset asthma. Although not identified in the previous SARP study, the Leicester clusters alluded to a similar subgroup, perhaps on the basis of inclusion of eosinophilic inflammation.³ Similarly, cluster 3 identifies a mixed group (>50% categorized as having SA) of primarily women with high symptom burden, early-onset disease, relatively normal lung function, and little inflammation. This group was also potentially identifiable in the Leicester clusters but is distinct from the female-predominant late-onset cluster 3 observed in the earlier SARP clinical clusters.

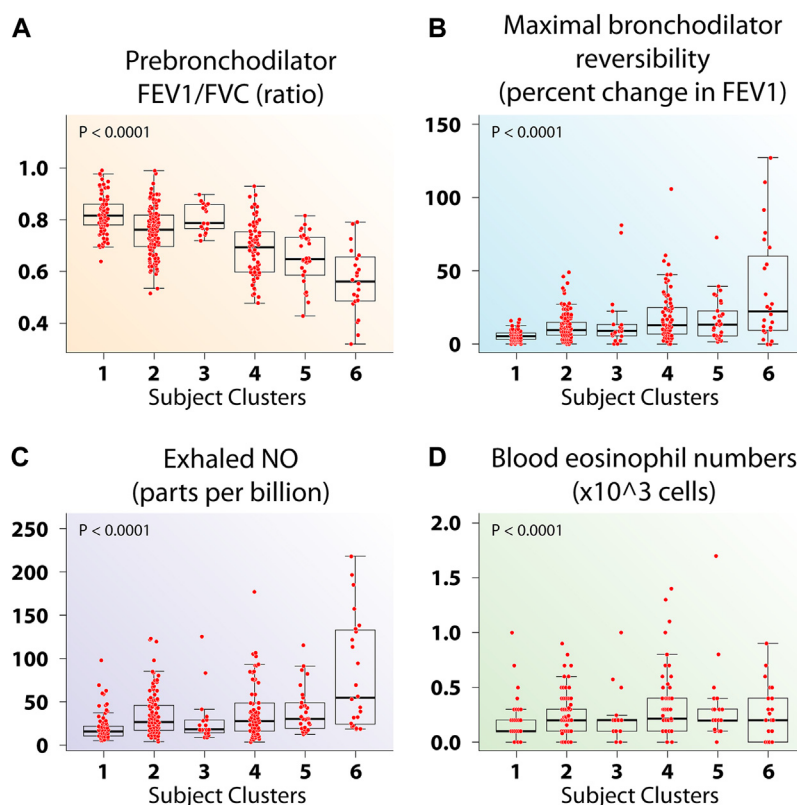


FIG 5. Airway responsiveness and T_H2 inflammatory markers differ across subject clusters. **A**, Prebronchodilator FEV₁/forced vital capacity (FVC) ratio was lower in cluster 6 than in all other clusters. **B**, There were no differences in reversibility across asthma clusters. **C**, Cluster 6 had higher FENO values compared with all other clusters. **D**, Clusters 2, 4, and 5 had higher blood eosinophil counts than cluster 1.

The presence of subject cluster 6 supports the importance of greater numbers of variables, even those with low INFOGAIN values. Our analysis suggests the nonredundant variables “use oral corticosteroids” or “systemic corticosteroids for $\geq 50\%$ of previous year,” which have reasonably high INFOGAIN values (0.33 and 0.28), separate subject cluster 6 from the other clusters, whereas blood neutrophil counts and osteoporosis, with low INFOGAIN values (0.12 and 0.1, respectively), are secondary. Interestingly, this hypothesis is consistent with the known effect of systemic corticosteroids to increase blood neutrophil counts and cause osteoporosis. Notably, despite systemic corticosteroids and their side effects, subject cluster 6 also has the highest FENO values, suggesting that the inflammation, morbidity, and HCU are not caused by poor compliance/adherence.

Subject cluster 4, which is characterized by early-onset, clinically poorly controlled allergic asthma (although only slightly $>50\%$ had predefined SA) with a prominent family history of asthma, differs markedly from cluster 3 (also $>50\%$ predefined SA). Cluster 4 has lower lung function than cluster 3. This cluster was identified in both the previous SARP and Leicester analyses.^{2,3}

This work was unique in that it used a suite of well-established machine learning techniques to analyze 112 variables. Both INFOGAIN and Markov blanket algorithms are machine learning approaches for feature selection. While INFOGAIN measurement allows quantification of how well each variable and cluster of variables distinguishes subject clusters, Markov

blanket algorithms find redundant variables in feature selection problems, which identify redundant variables for each nonredundant variable, being largely consistent with clinical observations. Thus machine learning–based analyses facilitated understanding of variable and subject clusters and enabled identification of the informative nonredundant distinguishing features for each subject cluster.

Our study is limited by the small sample size and must be tested and retested in larger independent data sets. Furthermore, the cross-sectional nature does not allow us to address cluster stability over time.

In summary, we clustered 112 variables into 10 categories and 378 SARP subjects into 6 groups. This analysis confirms the presence of previously identified clusters: (1) generally early-onset allergic asthma of a range of severity (all previous clusters); (2) a very severe, systemic corticosteroid-dependent, mixed inflammatory cluster (previous SARP analysis); (3) a late-onset eosinophilic cluster associated with nasal polyposis (Leicester clusters); and (4) a female cluster with symptom-predominant early-onset disease (Leicester clusters). Thus these results extend and bridge the original SARP and Leicester clusters. Whether this unsupervised approach to phenotyping will translate into an enhanced understanding of pathobiologic differences and guide therapy awaits further study.

We thank George Tseng at the Department of Biostatistics at the University of Pittsburgh for critical reading of the manuscript.

Key messages

- **Unsupervised analysis revealed several clinically recognizable patient clusters: an early-onset allergic asthma cluster; a very severe, systemic corticosteroid-dependent cluster; a late-onset nasal polyp cluster; and a cluster of women with symptom-predominant early-onset disease.**
- **Our results confirm, integrate, and expand on asthma phenotypes.**

REFERENCES

1. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 2012;18:716-25.
2. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 2010;181:315-23.
3. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008;178:218-24.
4. Siroux V, Garcia-Aymerich J. The investigation of asthma phenotypes. *Curr Opin Allergy Clin Immunol* 2011;11:393-9.
5. Fitzpatrick AM, Teague WG, Meyers DA, Peters SP, Li X, Li H, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *J Allergy Clin Immunol* 2011;127:382-9, e1-13.
6. 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. *J Allergy Clin Immunol* 2012;130:103-10.e8.
7. 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.
8. Gouvis-Echraghi R, Saint-Pierre P, Besharaty AA, Bernard A, Just J. Exhaled nitric oxide measurement confirms 2 severe wheeze phenotypes in young children from the Trousseau Asthma Program. *J Allergy Clin Immunol* 2012;130:1005-7.e1.
9. Koller D, Sahami M. Toward optimal feature selection. In: *Proceedings of the Thirteenth International Conference on Machine Learning*. Bari, Italy. July 3-6, 1996. pp. 284-92.
10. Xing EP, Jordan MI, Karp RM. Feature selection for high-dimensional genomic microarray data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. Williamstown, MA. June 28-July 1, 2001.
11. *Proceedings of the ATS workshop on refractory asthma: current understanding, recommendations, and unanswered questions*. American Thoracic Society. *Am J Respir Crit Care Med* 2000;162:2341-51.
12. Moore WC, Bleecker ER, Curran-Everett D, Erzurum SC, Ameredes BT, Bacharier L, et al. Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. *J Allergy Clin Immunol* 2007;119:405-13.
13. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;17:520-5.
14. Welch BL. The generalization of student's problems when several different population variances are involved. *Biometrika* 1947;34:28-35.
15. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002;12:111-39.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* 1995;57:289-300.
17. Mitchell TM. *Machine learning*. New York: McGraw-Hill Science; 1997.
18. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *J Machine Learning Res* 2004;5:1205-24.
19. Miranda C, Busacker A, Balzar S, Trudeau J, Wenzel SE. Distinguishing severe asthma phenotypes: role of age at onset and eosinophilic inflammation. *J Allergy Clin Immunol* 2004;113:101-8.
20. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010;363:1211-21.
21. Saffar AS, Ashdown H, Gounni AS. The molecular mechanisms of glucocorticoids-mediated neutrophil survival. *Curr Drug Targets* 2011;12:556-62.

Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data

Wei Wu, Eugene Bleecker, Wendy Moore, William Busse, Mario Castro, Kian Fan Chung, Serpil Erzurum, Ben Gaston, Elliot Israel, Douglas Curran-Everett, Sally E. Wenzel

Online Repository Materials

Methods:

Patient population

Severe asthma was defined per ATS recommendations.^{E1} Subjects are required to meet criteria for asthma and additionally meet 1 of 2 major criteria which 1) continual treatment with high dose inhaled corticosteroids (CS) and/or 2) systemic CSs for 50% or more of the previous year. In addition, 2 of 7 minor criteria are required including those related to daily symptoms, frequent exacerbations, severe exacerbations, FEV1<80% predicted, treatment with an additional controller agent, destabilization as CS therapy is reduced and history of intubation. Mild/no ICS had an FEV1% predicted of ≥80% on no ICS, Mild + ICS had an FEV1% predicted of ≥80% on low to moderate dose ICS, Moderate/no ICS had an FEV1% predicted of <80% on no ICS, while Moderate + ICS had an FEV1% predicted of <80% on low to moderate dose ICS. None of the controls met ICS dosing levels for severe asthma. Normal controls were without any respiratory symptoms and had normal lung function. No subject smoked >10 pack years and none had smoked in the previous year. Characterization included questionnaires on a range of topics, including symptom frequency, asthma specific quality of life (AQLQ, with increasing values signifying better AQLQ), medication use, health care utilization, family history, allergic history, and triggering factors. Pulmonary function testing was performed according to ATS guidelines as previously described.^{E2} Blood was collected for total serum IgE and a complete blood count and run by each site's clinical labs. F_ENO was measured at 50 mls/sec.^{E3} Atopy was addressed by skin prick testing to 12 common allergens as previously described.^{E2, 4} Bronchoscopy and lavage were performed per SARP protocol.^{E5} Following centrifugation to separate cells, cell counts and differentials were performed. All protocols were approved by the individual Institutional Regulatory Boards and subjects provided written informed consent. This 2nd clustering of the SARP data was limited to those subjects with "lung inflammatory" variables, including BAL cell counts and F_ENO.

Data Preprocessing:

Our dataset contained 38 continuous variables and 74 categorical variables, the latter of which included 71 ordinal and 3 nominal variables (“gender”, “ethnicity”, and “race”) (see Table E3 for details). To facilitate clustering of the subjects in these data, we created dummy variables for nominal variables, that is, for each nominal variable with M discrete values and $M \geq 2$, it was replaced by M binary indicator variables corresponding to each nominal value of the variable. Then to facilitate including the normal controls in the cluster analysis, missing values for “age asthma onset” and “asthma duration” in normal control subjects were entered as 90 (years old) and 0 (year), respectively. NCs all answered questions on symptoms, medications and health care utilization, such that the original values were utilized. It should be noted that occasionally normal controls reported some low levels of symptoms, despite otherwise being NCs. Utilizing different values for “age asthma onset” such as 200 did not change the results of the cluster analysis. Next, we removed all variables with data missing in $\geq 5\%$ of the subjects, while employing a k -NNimpute algorithm^{E6} to fill in values in those with $< 5\%$ missing data, which resulted in 78 variables with 699 data points imputed. The k -NNimpute algorithm first selects K subjects (among all 378 subjects) that are most similar to the subjects of interest, then imputes missing values in the subjects of interest based on the weighted average of the values of the variables associated with the K subjects. The k -NNimpute algorithm was implemented using the *impute.knn* function in the Bioconductor *impute* package with K chosen to be 5.^{E7}

Cluster analysis:

Clustering of the variables: An agglomerative hierarchical clustering with the Ward’s method^{E8} was applied to the preprocessed SARP data to partition the variables into distinctive groups. The nonparametric Spearman’s rank correlation coefficient was employed to measure similarities between variables. Since the hierarchical clustering algorithm used dissimilarity to measure distances between variables, similarity was converted to a dissimilarity measure by the following transformation: $d(v_i, v_j) = 1 - |R(v_i, v_j)|$, where $R(v_i, v_j)$ denotes similarity between variables v_i and v_j , and $d(v_i, v_j)$ denotes dissimilarity between v_i and v_j . The hierarchical clustering algorithm was implemented using the *hclust* function in the *stats* R package. Based on the dendrogram of the hierarchical tree of the variables resulting from the hierarchical clustering analysis and also additional clinical expert opinions (provided by Dr. Sally Wenzel), a height of 1.8 was used as a cutoff distance value to determine the number of the variable clusters.

Clustering of the subjects: Clustering of the subjects was performed independently of clustering of the variables. The preprocessed data was standardized so that each variable had mean 0 and variance 1. The k-means clustering method was applied to the standardized data to partition the subjects into distinctive groups. Let $S = \{S_1, S_2, \dots, S_N\}$ denote a set of subjects where $N = 378$ and each subject has M -dimensional clinical data where $M = 112$. Let $C = \{C_1, C_2, \dots, C_K\}$ denote K clusters with class labels from C_1 to C_K and K is preassigned. To partition the set S subjects into K clusters, the k-means algorithm searches for a partition which has minimal total within-cluster sum of squares^{E9}:

$$\arg \min_C \sum_{i=1}^K \sum_{x_j \in C_i} \|x_j - \mu_i\|,$$

where x_j is the data for subject j in cluster C_i , and μ_i is the mean of the subjects in cluster C_i . Euclidean distance was used to measure dissimilarities between different subjects. Even though the Gower distance was often used to measure dissimilarity for data with mixed types^{E10}, we found that Euclidean distance yields results which make better clinical sense. This agrees with recently published findings suggesting that Euclidean distance is a preferred dissimilarity measure for data with mixed types both theoretically and empirically.^{E11} A thousand iterations were performed to find an optimal partition of K clusters where K is pre-assigned. The k-means method is implemented based on the *kmeans* function in the *stats* R package with minor modification.

To partition the 378 subjects, we pre-assigned K to be from 4 to 7. We employed the average silhouette width (ASW)^{E12} as a criterion to determine an optimal number of clusters in the data. ASW is one of the most commonly used criteria for finding the number of clusters in the data. In order to calculate ASW, silhouettes of all objects in the data need to be computed. More specifically, a silhouette of an object i in a cluster is computed by comparing dissimilarities of object i with all other objects in the same cluster and dissimilarities of object i with all objects in the neighboring cluster. ASW is obtained by taking the mean of the silhouettes for all objects in the data. Silhouettes were calculated using the *silhouette* function in the *cluster* R package. We found that ASW was the largest when $K = 6$. In addition to using ASW as a statistical criterion to decide the cluster number, we also employed visualization and clinical criteria to validate our clustering results. When $K = 6$ was used to cluster the subjects, the k-means algorithm generated the clusters which not only could be clearly distinguished from one another using the heatmap (Figure 1), but were also clinically recognizable. We also performed cluster analysis with NCs removed completely from the data. The results remained mostly the same except that the two mild asthmatics in original subject Cluster 1 (Table 1) moved to Cluster 2.

Statistical tests:

Data for continuous variables were log transformed to improve normality of distribution. To find differences between N_T subjects in the total SARP cohort and the N subjects in this cluster analysis, we first employed Welch's t-tests^{E13}, which allow for data in different groups with unequal variances and are popularly used in microarray analysis^{E14}, to obtain t-statistics for the log2-transformed data from continuous variables. Then to account for the dependency structure in the total cohort and the one used in the cluster analysis, we employed a bootstrap procedure to estimate unadjusted p-values for the variables. The bootstrap procedure is as follows: N samples were randomly drawn from the N_T subjects in the total cohort, and a t-statistic t_b for a continuous variable v was calculated based on these samples. The same process was repeated $B = 10,000$ times to estimate an empirical distribution for the t-statistics of variable v . The bootstrap p-value p_v for the variable is obtained by the following formula: $p_v = \sum_{b=1}^B I(|t_b| \geq |T_v|)/B$, where $I(\cdot)$ is the indicator function and T_v denotes the t-statistic for variable v in the cohort we used for cluster analysis. The p-values for data from the categorical variables were obtained from Chi-square tests. P-values were adjusted for multiple testing using the FDR procedure. P-values < 0.05 were considered significant.

Discretization of data:

To calculate INFOGAIN, continuous variables need to be discretized. We used a simple unsupervised equal-width algorithm to discretize the data^{E15}. Specifically, given a variable X which has values ranging from x_{min} to x_{max} , X can be divided into K (the number of subject clusters) equal-width interval, and each interval has width $(x_{max} - x_{min})/K$. If a value x_i in X falls into the d th (d is an integer) interval of X , where

$$\frac{(x_{max}-x_{min})}{K} \cdot (d-1) \leq x_i \leq \frac{(x_{max}-x_{min})}{K} \cdot d,$$

x_i was discretized as d . Variables with categorical data which has the number of unique values greater than K were discretized using the same algorithm.

The requirement of data discretization for calculation of information gain (INFOGAIN) brings us a benefit, that is, the missing values in normal control (NC) subjects for "age asthma onset" and "asthma duration" can be represented by any distinct discrete values and the INFOGAINs for these two variables will not be affected.

Ranking variables using INFOGAIN:

We employed INFOGAIN^{E16} to measure how well a variable (X) predicts K subject cluster labels $L = \{L_1, L_2, \dots, L_K\}$ determined by the k-means clustering analysis. The INFOGAIN of X given L is defined as follows: Let $H(X)$ denote the entropy of X , and $H(X)$ is defined as $H(X) = -\sum_i P(x_i) \log_2(P(x_i))$, where $P(x_i)$ is the probability of value i in X . The entropy of X given L is defined as: $H(X|L) = -\sum_j P(l_j) \sum_i P(x_i|l_j) \log_2(P(x_i|l_j))$, where $P(l_j)$ is the probability of value j in L , and $P(x_i|l_j)$ is the probability of value i in X given the value j in L . The INFOGAIN of X given L is defined as: $INFOGAIN(X|L) = H(X) - H(X|L)$.

Selection of relevant and nonredundant variables:

Given an initial set of features $F = \{F_1, F_2, \dots, F_n\}$ and K subject cluster labels $L = \{L_1, L_2, \dots, L_k\}$, the features in F were sorted, so that:

$$INFOGAIN(F_1, L) \geq INFOGAIN(F_2, L) \geq \dots \geq INFOGAIN(F_n, L).$$

Then, we selected a set of relevant features F_{rel} from F , which were defined as features each correlated with the cluster labels L above a threshold value δ and $\delta = 0.05$, and thus providing sufficient information for L . Here, we used INFOGAIN to correlate feature F_i with cluster labels L . If $INFOGAIN(F_i, L) \geq 0.05$, then we considered F_i as relevant. Then the Markov-blanket algorithm^{E17} selects nonredundant features in F_{rel} as follows. For any feature F_j in F_{rel} , a feature F_m , where $m \neq j$, is considered redundant if F_j is an approximate Markov Blanket for F_m in F_{rel} , if and only if the following condition is met: $INFOGAIN(F_j, L) \geq INFOGAIN(F_m, L)$, and $INFOGAIN(F_j, F_m) \geq INFOGAIN(F_m, L)$. The redundant feature F_m for F_j was eliminated from the feature set F_{rel} ; this process was repeated until all the redundant features for F_j were removed and the new feature set F'_{rel} becomes: $F'_{rel} = F_{rel} - \{\text{redundant features for } F_j\}$. The same process was repeated until we reached the last feature in the final feature set F_{final} , and F_{final} contains all the relevant, nonredundant features for the initial feature set F . To favor selection of variables which give better distinction for multi-class rather than binary-class data, we used INFOGAIN, rather than normalized INFOGAIN, in the original algorithm to measure similarity between variables.

Classification of clustered subjects using a Support Vector Machine (SVM) algorithm:

SVMs are among the most effective supervised learning techniques^{E18}, and are robust to the overfitting problem often associated with classification using high dimensional data. They have been applied to many areas of

biological analysis with superior classification performance, including biomarker identification and cancer classification.^{E19-21} In order to classify samples with binary classes, SVMs first identify a hyperplane with a maximum margin that separates different classes of the training samples. In cases when samples of different classes are nonseparable, data is mapped by a kernel function into a high-dimensional space where data becomes separable by a hyperplane. Then, new samples are classified according to the side of the hyperplane they belong to.

In order to classify subjects whose cluster labels were assigned by k-means clustering, we employed a multi-class SVM algorithm with a linear kernel which utilized a ‘one-versus-one’ strategy^{E22} to classify samples with multiple classes. More specifically, let K denote the number of classes and $K > 2$, $K(K - 1)/2$ binary SVM classifiers were first constructed for all pairs of the classes with the training samples. A new sample was assigned a class which had the largest number of the votes from all the classifiers. The SVM algorithm is implemented based on the *svm* function in the *e1071* R package. The 378 subjects were split into a training data (80% subjects) and a test data (20% subjects). The SVM predication on the test data, and also the prediction on the training data by leave-one-out cross validation (*LOOCV*) was performed to estimate classification errors. This procedure was repeated 10 times to ensure an accurate estimate of classification error rates.

Results:

Predictive values of the variables

Classification with all the variables

Using a multi-class SVM classifier with all the variables, we performed prediction on the training data with *LOOCV* as well as prediction on the test data. On average, the accuracy rate of SVM classification for *LOOCV* on the training data is 93% ($\pm 1\%$), and the accuracy rate of prediction on the test data is 93% ($\pm 2\%$).

Classification with selected relevant, nonredundant variables

Since many variables in the SARP data are highly correlated with one another, the impact of removing redundant variables was addressed, employing a machine learning procedure for feature selection and exploring how well subject cluster labels were predicted after removing redundant variables. In each set of the training data, the relevant variables were selected and a list of relevant, nonredundant variables were identified using a Markov blanket algorithm. Since 10 sets of training and test data were generated and used for our variable selection and classification tests, slightly different sets of relevant nonredundant variables were obtained from different training

data sets. To illustrate this variable selection process, the Markov-blanket algorithm was applied to the full SARP dataset (with the 378 subjects). Table E3 shows the 51 relevant nonredundant variables selected as well as the redundant variables identified with respect to each nonredundant variable from the whole dataset. Finally, the predictive value of the cluster labels was addressed using the selected relevant nonredundant features. On average, the SVM classification accuracy rate for *LOOCV* on the training data is 92% ($\pm 2\%$), and the accuracy rate of SVM prediction on the test data is 88% ($\pm 3\%$).

Discussion:

Markov blanket algorithms find redundant variables in feature selection problems. Intuitively, the algorithm identifies a redundant variable X as follows. If Y contains more information about the class labels L than X , then given Y , X is considered redundant for predicting class labels L .

References:

- E1. Proceedings of the ATS workshop on refractory asthma: current understanding, recommendations, and unanswered questions. American Thoracic Society. Am J Respir Crit Care Med 2000; 162:2341-51.
- E2. Moore WC, Bleecker ER, Curran-Everett D, Erzurum SC, Ameredes BT, Bacharier L, et al. Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. J Allergy Clin Immunol 2007; 119:405-13.
- E3. ATS/ERS recommendations for standardized procedures for the online and offline measurement of exhaled lower respiratory nitric oxide and nasal nitric oxide, 2005. Am J Respir Crit Care Med 2005; 171:912-30.
- E4. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med 2010; 181:315-23.
- E5. Moore WC, Evans MD, Bleecker ER, Busse WW, Calhoun WJ, Castro M, et al. Safety of investigative bronchoscopy in the Severe Asthma Research Program. J Allergy Clin Immunol 2011; 128:328-36 e3.
- E6. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17:520-5.

196 E7. Hastie T, Tibshirani R, Narasimhan B, Chu G. The "impute" R package; 2012. Available from
 197 <http://www.bioconductor.org/packages/release/bioc/html/impute.html>.

198 E8. Ward JH. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical
 199 Association 1963; 58:236–44.

200 E9. Hartigan JA, Wong MA. A K-means clustering algorithm. Applied Statistics 1979; 28:100–8.

201 E10. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics 1971; 27:857-71.

202 E11. Hennig C, Liao TF. How to find an appropriate clustering for mixed type variables with application to
 203 socio-economic stratification. Applied Statistics 2013; 62:1–25.

204 E12. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis: Wiley-
 205 Interscience; 1990.

206 E13. Welch BL. The generalization of "Student's" problem when several different population variances are
 207 involved". Biometrika 1947; 34:28–35.

208 E14. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed
 209 genes in replicated cDNA microarray experiments. Statistica Sinica 2002; 12:111-39.

210 E15. Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of Continuous Features.
 211 Proceedings of the Twelfth International Conference on Machine Learning 1995:194--202.

212 E16. Mitchell TM. Machine Learning: McGraw-Hill Science; 1997.

213 E17. Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine
 214 Learning Research 2004; 5:1205–24.

215 E18. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995.

216 E19. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine
 217 classification and validation of cancer tissue samples using microarray expression data. Bioinformatics
 218 2000; 16:906-14.

219 E20. Zhang F, Kaufman HL, Deng Y, Drabier R. Recursive SVM biomarker selection for early detection of
 220 breast cancer in peripheral blood. BMC Med Genomics 2013; 6 Suppl 1:S4.

221 E21. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory
 222 classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005; 21:631-43.

223 E22. Kressel U. Pairwise classification and support vector machines. In: Advances in Kernel Methods: Support
224 Vector Learning. Cambridge, MA, USA: MIT Press; 1999.
225
226

Table E1: Demographic data of the subjects in the SARP cohort.

	Subjects in the SARP cohort	Subjects for Cluster Analysis	P-value* (FDR)
Age at Enrollment	33 ± 15 [†]	35 ± 12	< 0.0001
% Male/female	633(39%)/1024 (61%)	142 (38%)/236 (62%)	0.86
% Caucasians/AA/Other	1021 (62%)/498 (30%)/166 (8%)	270 (71%)/78 (21%)/30 (8%)	0.032
% Hispanic/Non-Hispanic/Unknown	71 (4%)/1559 (94%)/23 (2%)	19 (5%)/350 (93%)/8 (2%)	0.79
BMI	28 ± 8	28 ± 7	0.27
Prebronchodilator FEV1 (% predicted)	81 ± 22	83 ± 20	0.0017
Severity Levels (n) [#]			0.058
Normals	283 (17%)	89 (23%)	
Mild/No ICS	214 (13%)	57 (15%)	
Mild + ICS	293 (17%)	63 (17%)	
Moderate/no ICS	107 (6%)	20 (5%)	
Moderate + ICS	169 (10%)	37 (10%)	
Severe	589 (35%)	112 (30%)	
Total	1685	378	

* FDR adjusted *p*-value from Welch's *t*-tests or Chi-square tests between the 1685 subjects in the total SARP cohort and the 378 subjects used for the cluster analysis

[†] Numerical data are presented as antilog of log2-transformed mean ± SD.

[#] See methods for detailed description

Table E2: Ten clusters of variables discovered by cluster analysis. For each cluster of variables (listed under the “Variables” column), those shown in the upper row have INFOGAINs $\geq$ the median INFOGAIN of the cluster as indicated in the third column, while those in the lower row have INFOGAINs $<$ the median INFOGAIN. Variables in each cluster are ordered by the magnitude of their INFOGAIN, and the one with the largest or the lowest INFOGAIN is shown in the parenthesis.

Cluster	Variables	Median INFOGAIN
1	Shortness of breath frequency (0.81), Activity, Symptoms, Wheeze, Chest tightness	0.68
	Emotion, Environment, Night-time awakenings, Cough frequency, Sputum frequency (0.31)	
2	Age asthma onset (0.84), Asthma medication use in last 3 mos, Asthma worsened by upper respiratory infection, Use inhaled β -agonist, Asthma duration, Allergy symptoms in winter, Allergy symptoms in fall	0.40
	Asthma symptoms caused by exercise, Allergy symptoms in spring, Allergy symptoms in summer, Asthma symptoms caused by animal exposure, Had allergy symptoms without cold or flu, Number of skin reactions to allergens, atopy, IgE level (0.01)	
3	Prebronchodilator FEV1 (0.48), Maximum post bronchodilator FEV1	0.29
	Prebronchodilator FVC, Maximum post bronchodilator FVC (0.12)	
4	Seen MD in last 12 mos for asthma (0.48), Take high dose inhaled CS, Asthma symptoms caused by activity, ER visit in last yr, >3 oral CS bursts in previous year, Use daily β -agonist, Use oral CSs, Use combination therapy, Symptoms worsen when CS decreased, Take systemic CS for 50% or more of last year, Have history of nasal polyps, Had surgery to remove nasal polyps, Use nebulized β -agonist, Visited ER for breathing in last 12 mos	0.21
	Use leukotriene modifier, Number of ICU admissions for asthma, Hospitalized for asthma in last 12 mos, Ever had acute or recurrent sinusitis, Ever had sinus surgery, Use nasal steroids, Use inhaled CSs, Use theophylline, Diagnosed with osteoporosis, Use ipratropium, Use anti-IgE therapy, Asthma worsened by nonsteroidal pain medications, Use long acting β -agonist, Ever intubated (0.06)	
5	Prebronchodilator FEV1/FVC (0.33), Maximum post bronchodilator FEV/FVC, Maximal bronchodilator reversibility, Exhaled NO	0.16
	Blood eosinophil percentage, Blood eosinophil numbers, BAL eosinophil percentage (0.09)	
6	Total WBC (0.19), Blood neutrophil percentage	0.12
	Blood neutrophil numbers, Blood lymphocyte percentage (0.12)	
7	Age (0.22), Ever had recurrent bronchitis, BMI, Number of children, Systolic BP, Ever had pneumonia	0.11
	Have GERD, Diastolic BP, Diagnosed with chronic bronchitis, Ever smoke, Smoking status [†] (0.02)	
8	BAL lymphocyte percentage [†] (0.08), BAL alveolar macrophage percentage	0.07
	BAL neutrophil percentage [†] (0.06)	
9	Height (0.11), Hematocrit	0.04
	Gender_Male, Gender_Female (0.037)	
10	Ethnicity_NonHispanic (0.23), Ethnicity_Hispanic, Mother/Father with asthma, Blood lymphocyte numbers, Ethnicity_Unknown, Number of siblings, Use injectable CSs, Exposed to 2nd hand smoke, Race_AfricanAmerican, Mother with asthma, Race_Caucasian, Blood monocyte numbers [†] , Blood monocyte percentage, Blood basophil numbers, Sibling with asthma, Had prolonged skin rash, Father with asthma	0.04
	Race_Multiracial, Diagnosed with diabetes, Father with allergy, Use alternative therapy, Race_Other, Use immunotherapy, Use combination albuterol-ipratropium, Have a cat [†] , Mother with allergy [†] , Mother with COPD [†] , Race_Asian [†] , Total cell count in BAL, Have a dog [†] , Have a bird [†] , Ethnicity_Refused [†] , Father with eczema [†] , Mother with eczema [†] (0.002)	

[†] shows the variables which did not differ significantly among the subject clusters, all the other variables were significantly different among the subject clusters.

Table E3. The 112 variables used in the cluster analysis and their INFOGAIN values. *Variables are ordered by their INFOGAIN values, in a descending order. Variable types are listed in the parenthesis following the names of the variables: “Con” stands for continuous variable; “Cat, ord” stands for categorical ordinal variable; and “Cat, bin” stands for categorical binary variable. † FDR adjusted *p*-value from ANOVA, Kruskal-Wallis rank sum test, or Chi-square analysis for continuous, categorical ordinal and categorical binary variables, respectively, between six subject clusters. ‡ Clusters with significant differences determined by pairwise *t*-test, Wilcoxon rank sum test, or Fisher's exact test for continuous, categorical ordinal and categorical binary variables (FDR adjusted *p*-value < 0.05).

Variable*	INFOGAIN	P-value (FDR) [†]	Significantly Different Clusters (FDR <i>p</i> -value< 0.05) [‡]
1. Age asthma onset (Con)	0.84	1.20E-53	Cl2-Cl5; Cl3-Cl5; Cl4-Cl5; Cl5-Cl6
2. Shortness of breath frequency (Cat, ord)	0.81	1.40E-48	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
3. Activity (Con)	0.78	5.00E-71	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
4. Symptoms (Con)	0.74	1.20E-69	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
5. Wheeze (Cat, ord)	0.72	6.50E-45	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
6. Chest tightness (Cat, ord)	0.71	9.70E-44	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
7. Asthma medication use in last 3 mos (Cat, bin)	0.66	4.40E-64	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
8. Emotion (Con)	0.65	4.60E-59	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
9. Environment (Con)	0.61	7.70E-47	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl6; Cl4-Cl5; Cl5-Cl6
10. Night-time awakenings (Cat, ord)	0.57	1.50E-39	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
11. Asthma worsened by upper respiratory infection (Cat, bin)	0.56	1.30E-58	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
12. Use inhaled beta-agonist (Cat, bin)	0.55	9.80E-53	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
13. Asthma duration (Con)	0.49	8.30E-139	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl5-Cl6
14. Seen MD in last 12 mos for asthma (Cat, bin)	0.48	3.70E-40	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6
15. Prebronchodilator FEV1 (Con)	0.48	1.40E-52	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
16. Cough frequency (Cat, ord)	0.47	4.90E-34	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
17. Allergy symptoms in winter (Cat, ord)	0.42	1.20E-29	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6
18. Allergy symptoms in fall (Cat, ord)	0.41	6.80E-29	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl6; Cl5-Cl6
19. Take high dose inhaled CS (Cat, bin)	0.41	3.40E-38	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6
20. Asthma symptoms caused by exercise (Cat, bin)	0.4	4.70E-40	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
21. Asthma symptoms caused by activity (Cat, bin)	0.39	1.80E-34	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6
22. ER visit in last yr (Cat, bin)	0.36	6.10E-34	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
23. >3 oral CS bursts in previous year (Cat, bin)	0.35	7.90E-37	Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
24. Use daily beta-agonist (Cat, bin)	0.34	9.00E-31	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6
25. Allergy symptoms in spring (Cat, ord)	0.34	6.60E-25	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
26. Maximum post bronchodilator FEV1 (Con)	0.34	5.30E-35	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
27. Prebronchodilator FEV1/FVC (Con)	0.33	1.30E-38	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
28. Use oral CSs (Cat, bin)	0.33	6.00E-41	Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
29. Use combination therapy (Cat, bin)	0.32	8.40E-26	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6
30. Sputum frequency (Cat, ord)	0.31	4.30E-23	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-

			Cl6
31. Symptoms worsen when CS decreased (Cat, bin)	0.3	1.90E-25	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6
32. Allergy symptoms in summer (Cat, ord)	0.3	6.90E-22	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl6
33. Asthma symptoms caused by animal exposure (Cat, bin)	0.3	2.40E-25	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
34. Take systemic CS for 50% or more of last year (Cat, bin)	0.28	2.20E-41	Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
35. Maximum post bronchodilator FEV/FVC (Con)	0.28	1.00E-27	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
36. Have history of nasal polyps (Cat, bin)	0.27	2.30E-46	Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
37. Had surgery to remove nasal polyps (Cat, bin)	0.26	9.40E-54	Cl1-Cl5; Cl2-Cl5; Cl3-Cl5; Cl4-Cl5; Cl5-Cl6
38. Had allergy symptoms without cold or flu (Cat, bin)	0.25	2.30E-28	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
39. Use nebulized beta-agonist (Cat, bin)	0.25	5.40E-26	Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
40. Prebronchodilator FVC (Con)	0.24	1.80E-27	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6; Cl4-Cl6
41. Ethnicity_NonHispanic (Cat, bin)	0.23	1.80E-51	Cl1-Cl3; Cl2-Cl3; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6
42. Age (Con)	0.22	2.90E-17	Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl3-Cl6; Cl4-Cl5
43. Number of skin reactions to allergens (Con)	0.22	1.60E-22	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl3-Cl5; Cl4-Cl5
44. Visited ER for breathing in last 12 mos (Cat, bin)	0.21	3.30E-22	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
45. Use leukotriene modifier (Cat, bin)	0.21	6.50E-20	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
46. Atopy (Cat, bin)	0.19	8.50E-22	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
47. Total WBC (Con)	0.19	2.10E-17	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
48. Maximal bronchodilator reversibility (Con)	0.18	1.40E-11	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6
49. Number of ICU admissions for asthma (Cat, ord)	0.17	6.70E-15	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
50. Exhaled NO (Con)	0.16	1.30E-13	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
51. Hospitalized for asthma in last 12 mos (Cat, bin)	0.16	4.60E-24	Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
52. Ever had acute or recurrent sinusitis (Cat, bin)	0.16	6.50E-15	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl2-Cl6; Cl3-Cl5; Cl4-Cl5; Cl4-Cl6
53. Ever had recurrent bronchitis (Cat, bin)	0.15	2.90E-14	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6
54. Ever had sinus surgery (Cat, bin)	0.15	4.60E-22	Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl3-Cl5; Cl4-Cl5; Cl5-Cl6
55. Use nasal steroids (Cat, bin)	0.14	9.10E-14	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6
56. BMI (Con)	0.14	1.80E-07	Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5
57. Use inhaled CSs (Cat, bin)	0.13	2.90E-12	Cl1-Cl2; Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6; Cl5-Cl6
58. Blood neutrophil percentage (Con)	0.12	5.60E-07	Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6
59. Ethnicity_Hispanic (Cat, bin)	0.12	1.00E-28	Cl1-Cl3; Cl2-Cl3; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6
60. Blood neutrophil numbers (Con)	0.12	6.90E-13	Cl1-Cl2; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
61. Blood lymphocyte percentage (Con)	0.12	4.20E-14	Cl1-Cl5; Cl1-Cl6; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6
62. Number of children (Cat, ord)	0.12	1.60E-11	Cl1-Cl3; Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6; Cl4-Cl6
63. Maximum post bronchodilator FVC (Con)	0.12	9.60E-10	Cl1-Cl4; Cl1-Cl5; Cl1-Cl6; Cl2-Cl4; Cl2-Cl5; Cl2-Cl6; Cl3-Cl6
64. Systolic BP (Con)	0.12	2.00E-07	Cl1-Cl4; Cl1-Cl5; Cl2-Cl3; Cl2-Cl4; Cl2-Cl5; Cl3-Cl4; Cl3-Cl5; Cl3-Cl6
65. Height (Con)	0.11	6.60E-06	Cl1-Cl3; Cl1-Cl6; Cl2-Cl3; Cl2-Cl4; Cl2-Cl6; Cl3-Cl4; Cl3-Cl5; Cl5-Cl6
66. Ever had pneumonia (Cat, bin)	0.11	4.00E-11	Cl1-Cl4; Cl1-Cl6; Cl2-Cl4; Cl2-Cl6; Cl3-Cl6; Cl4-Cl5; Cl4-Cl6; Cl5-Cl6
67. Use theophylline (Cat, bin)	0.11	1.80E-18	Cl1-Cl3; Cl1-Cl6; Cl2-Cl3; Cl2-Cl6; Cl4-Cl6; Cl5-Cl6

68. Have GERD (Cat, bin)	0.1	1.30E-10	C11-C14; C11-C15; C11-C16; C12-C14; C12-C15; C12-C16; C13-C16
69. Diagnosed with osteoporosis (Cat, bin)	0.1	1.30E-18	C11-C16; C12-C15; C12-C16; C13-C16; C14-C16; C15-C16
70. Diastolic BP (Con)	0.1	1.50E-04	C11-C14; C11-C15; C12-C14; C12-C15; C13-C15
71. Blood eosinophil percentage (Con)	0.1	4.00E-05	C11-C12; C11-C14
72. Blood eosinophil numbers (Con)	0.09	8.40E-07	C11-C12; C11-C14; C11-C15
73. Mother/Father with asthma (Cat, bin)	0.09	3.90E-08	C11-C12; C11-C13; C11-C14; C11-C16; C12-C14; C14-C15
74. Blood lymphocyte numbers (Con)	0.09	9.50E-03	C11-C14
75. Use ipratropium (Cat, bin)	0.09	4.40E-16	C11-C16; C12-C16; C14-C16; C15-C16
76. BAL eosinophil percentage (Con)	0.09	7.60E-09	C11-C12; C11-C13; C11-C14; C11-C15; C11-C16; C12-C15; C13-C15; C14-C15
77. Ethnicity_Unknown (Cat, bin)	0.08	2.00E-23	C11-C13; C12-C13; C13-C14; C13-C15; C13-C16
78. BAL lymphocyte percentage (Con)	0.08	N.S. [§]	
79. Use anti-IgE therapy (Cat, bin)	0.08	1.60E-12	C11-C15; C11-C16; C12-C15; C12-C16; C13-C16; C14-C16
80. Number of siblings (Con)	0.08	9.10E-04	C12-C14; C12-C15
81. Use injectable CSs (Cat, bin)	0.08	5.90E-16	C11-C16; C12-C16; C14-C16
82. Exposed to 2nd hand smoke (Cat, bin)	0.07	1.80E-07	C11-C14; C12-C14; C14-C15
83. BAL alveolar macrophage percentage (Con)	0.07	1.40E-05	C11-C13; C13-C14; C13-C15; C13-C16
84. Asthma worsened by nonsteroidal pain medications (Cat, bin)	0.07	1.00E-08	C11-C15; C11-C16; C12-C15; C12-C16; C14-C15
85. Race_AfricanAmerican (Cat, bin)	0.06	3.40E-06	C11-C14; C11-C16; C12-C14; C13-C14; C14-C15; C15-C16
86. Mother with asthma (Cat, bin)	0.06	1.00E-05	C11-C12; C11-C14; C14-C15
87. Race_Caucasian (Cat, bin)	0.06	5.70E-06	C11-C13; C11-C14; C12-C13; C12-C14; C13-C15; C14-C15; C15-C16
88. Blood monocyte numbers (Con)	0.06	N.S.	
89. Diagnosed with chronic bronchitis (Cat, bin)	0.06	1.60E-05	C11-C14; C11-C15; C11-C16; C12-C14; C12-C16
90. BAL neutrophil percentage (Con)	0.06	N.S.	
91. Use long acting beta-agonist (Cat, bin)	0.06	1.60E-05	C11-C14; C11-C15; C11-C16; C12-C16
92. Blood monocyte percentage (Con)	0.06	1.60E-05	C11-C13; C11-C14; C11-C15; C11-C16
93. Ever intubated (Cat, bin)	0.06	7.40E-07	C11-C16; C12-C16; C14-C16
94. Blood basophil numbers (Con)	0.05	5.40E-03	C11-C14
95. Sibling with asthma (Cat, bin)	0.05	4.00E-04	C11-C14
96. Had prolonged skin rash (Cat, bin)	0.05	1.10E-03	C11-C12; C11-C14; C11-C15; C11-C16
97. Father with asthma (Cat, bin)	0.04	1.20E-03	C11-C13; C11-C14; C11-C16
98. Hematocrit (Con)	0.04	3.00E-02	C12-C15; C13-C15; C14-C15
99. Gender_Male (Cat, bin)	0.04	2.30E-03	C13-C15; C14-C15
100. Gender_Female (Cat, bin)	0.04	2.30E-03	C13-C15; C14-C15
101. Race_Multiracial (Cat, bin)	0.03	5.00E-04	C12-C13
102. Diagnosed with diabetes (Cat, bin)	0.03	7.60E-04	C11-C16; C12-C16
103. Father with allergy (Cat, bin)	0.03	1.30E-02	C11-C12
104. Use alternative therapy (Cat, bin)	0.03	1.40E-02	
105. Ever smoke (Cat, bin)	0.03	1.20E-02	C11-C14
106. Race_Other (Cat, bin)	0.03	8.20E-05	
107. Use immunotherapy (Cat, bin)	0.02	1.60E-04	
108. Use combination albuterol-ipratropium (Cat, bin)	0.02	6.00E-03	
109. Have a cat (Cat, bin)	0.02	N.S.	
110. Mother with allergy (Cat, bin)	0.02	N.S.	
111. Mother with COPD (Cat, bin)	0.02	N.S.	
112. Race_Asian (Cat, bin)	0.02	N.S.	
113. Smoking status (Cat, bin)	0.02	N.S.	
114. IgE level (Con)	0.01	3.60E-12	C11-C12; C11-C13; C11-C14; C11-C15
115. Total cell count in BAL (Con)	0.01	6.30E-06	C11-C16; C12-C16; C13-C16; C14-C16; C15-C16
116. Have a dog (Cat, bin)	0.01	N.S.	
117. Have a bird (Cat, bin)	0.01	N.S.	
118. Ethnicity_Refused (Cat, bin)	0.01	N.S.	
119. Father with eczema (Cat, bin)	0	N.S.	
120. Mother with eczema (Cat, bin)	0	N.S.	

[§]N.S. stands for “not significant”.

Table E4. The 51 relevant and nonredundant variables. The relevant and nonredundant variables are shown in the first column, while their corresponding redundant variables are listed in the second column.

Relevant and Nonredundant Variables	Redundant Variables
Age asthma onset	Asthma duration, Allergy symptoms in spring, Asthma symptoms caused by animal exposure, Had allergy symptoms without cold or flu, Age, Number of skin reactions to allergens, Atopy, BMI, Blood eosinophil percentage, Mother/Father with asthma
Shortness of breath frequency	Wheeze, Chest tightness, Night-time awakenings, Number of children
Activity	Symptoms, Emotion, Environment, Diagnosed with chronic bronchitis
Asthma medication use in last 3 mos Asthma worsened by upper respiratory infection Use inhaled β -agonist Seen MD in last 12 mos for asthma	NULL
Prebronchodilator FEV1	Maximum post bronchodilator FEV1, Prebronchodilator FEV1/FVC, Maximum post bronchodilator FEV/FVC, Prebronchodilator FVC, Maximal bronchodilator reversibility, Maximum post bronchodilator FVC, Diastolic BP, Blood eosinophil numbers, Number of siblings
Cough frequency	Sputum frequency
Allergy symptoms in winter	Allergy symptoms in fall, Allergy symptoms in summer, BAL neutrophil percentage
Take high dose inhaled CS Asthma symptoms caused by exercise Asthma symptoms caused by activity	NULL
ER visit in last yr	Visited ER for breathing in last 12 mos
>3 oral CS bursts in previous year Use daily β -agonist Use oral CSs Use combination therapy Symptoms worsen when CS decreased Take systemic CS for 50% or more of last year Have history of nasal polyps	NULL
Had surgery to remove nasal polyps	Ever had sinus surgery
Use nebulized β -agonist	NULL
Ethnicity_NonHispanic	Ethnicity_Hispanic
Use leukotriene modifier	NULL
Total WBC	Blood neutrophil percentage, Blood neutrophil numbers, Blood lymphocyte percentage, Blood lymphocyte numbers, Blood monocyte numbers, Blood monocyte percentage
Number of ICU admissions for asthma	Ever intubated
Exhaled NO Hospitalized for asthma in last 12 mos Ever had acute or recurrent sinusitis Ever had recurrent bronchitis Use nasal steroids Use inhaled CSs Systolic BP Height Ever had pneumonia Use theophylline Have GERD Diagnosed with osteoporosis Use ipratropium	NULL
BAL eosinophil percentage	BAL alveolar macrophage percentage
Ethnicity_Unknown BAL lymphocyte percentage Use anti-IgE therapy Use injectable CSs Exposed to 2nd hand smoke Asthma worsened by nonsteroidal pain medications	NULL
Race_AfricanAmerican	Race_Caucasian
Mother with asthma Use long acting β -agonist Blood basophil numbers	NULL

Table E5. Comparison of subject cluster labels in the current study with previous SARP analysis. For the 250 overlapping subjects, the left column shows the number of subjects in each cluster from the current study, while the right column summarizes how the participants in the current cluster are distributed in the previous SARP clusters. The numbers of subjects belonging to the clusters in either current or previous SARP study are displayed in the parenthesis after the cluster labels.

Current study	Previous SARP analysis
Subject cluster 1 (2)	Cluster 1 (1) + Cluster 2 (1)
Subject cluster 2 (114)	Cluster 1 (37) + Cluster 2 (61) + Cluster 3 (5) + Cluster 4 (9) + Cluster 5 (2)
Subject cluster 3 (17)	Cluster 1 (7) + Cluster 2 (8) + Cluster 4 (1) + Cluster 5 (1)
Subject cluster 4 (72)	Cluster 1 (2) + Cluster 2 (38) + Cluster 3 (9) + Cluster 4 (16) + Cluster 5 (7)
Subject cluster 5 (23)	Cluster 1 (1) + Cluster 2 (7) + Cluster 3 (7) + Cluster 4 (5) + Cluster 5 (3)
Subject cluster 6 (22)	Cluster 1 (1) + Cluster 3 (1) + Cluster 4 (14) + Cluster 5 (6)

Table E6. Summary of the data from the 112 variables in the subject clusters. *Variables are in the same order as shown in Table E3. [†] FDR adjusted *p*-values are also the same as shown in Table E3. [‡] Numerical data are presented as mean ± SD.

Variable*	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	P-value (FDR) [†]
Age asthma onset (age)	NULL	11 +/- 10 [‡]	15 +/- 13	14 +/- 13	29 +/- 17	14 +/- 15	1.20E-53
Shortness of breath frequency (score)	1 +/- 0	3 +/- 1	4 +/- 1	4 +/- 1	4 +/- 1	5 +/- 1	1.40E-48
Activity (score)	7 +/- 0	6 +/- 1	4 +/- 1	4 +/- 1	4 +/- 1	3 +/- 1	5.00E-71
Symptoms (score)	7 +/- 0	6 +/- 1	5 +/- 1	4 +/- 1	5 +/- 1	3 +/- 1	1.20E-69
Wheeze (score)	1 +/- 0	2 +/- 1	3 +/- 1	4 +/- 1	4 +/- 1	5 +/- 1	6.50E-45
Chest tightness (score)	1 +/- 0	3 +/- 1	3 +/- 1	4 +/- 1	3 +/- 1	5 +/- 1	9.70E-44
Asthma medication use in last 3 mos (%)	100/0	8/92	5/95	1/99	0/100	4/96	4.40E-64
Emotion (score)	7 +/- 0	6 +/- 1	4 +/- 2	3 +/- 1	5 +/- 2	3 +/- 1	4.60E-59
Environment (score)	7 +/- 0	5 +/- 1	4 +/- 1	4 +/- 1	5 +/- 2	3 +/- 1	7.70E-47
Night-time awakenings (score)	1 +/- 0	2 +/- 1	3 +/- 1	3 +/- 1	3 +/- 1	5 +/- 1	1.50E-39
Asthma worsened by upper respiratory infection (%)	95/5	7/93	5/95	1/99	7/93	0/100	1.30E-58
Use inhaled beta-agonist (%)	100/0	10/90	5/95	10/90	17/83	8/92	9.80E-53
Asthma duration (yrs)	0	21 +/- 11	19 +/- 10	24 +/- 12	18 +/- 12	29 +/- 15	8.30E-139
Seen MD in last 12 mos for asthma (%)	100/0	64/36	16/84	14/86	3/97	4/96	3.70E-40
Prebronchodilator FEV1 (% predicted)	100 +/- 11	88 +/- 15	93 +/- 13	71 +/- 14	65 +/- 17	50 +/- 21	1.40E-52
Cough frequency (score)	1 +/- 1	2 +/- 1	3 +/- 1	4 +/- 2	3 +/- 2	4 +/- 2	4.90E-34
Allergy symptoms in winter (score)	1 +/- 0	2 +/- 1	2 +/- 1	3 +/- 1	2 +/- 1	3 +/- 1	1.20E-29
Allergy symptoms in fall (score)	1 +/- 1	2 +/- 1	3 +/- 1	3 +/- 1	3 +/- 1	3 +/- 1	6.80E-29
Take high dose inhaled CS (%)	100/0	92/8	47/53	44/56	13/87	8/92	3.40E-38
Asthma symptoms caused by exercise (score)	92/8	16/84	11/89	16/84	10/90	0/100	4.70E-40
Asthma symptoms caused by activity (%)	98/2	64/36	21/79	16/84	13/87	4/96	1.80E-34
ER visit in last yr (%)	100/0	90/10	74/26	42/58	43/57	0/100	6.10E-34
>3 oral CS bursts in previous year (%)	100/0	99/1	79/21	70/30	43/57	4/96	7.90E-37
Use daily beta-agonist (%)	100/0	83/17	42/58	30/70	40/60	12/88	9.00E-31
Allergy symptoms in spring (score)	1 +/- 1	3 +/- 1	3 +/- 1	3 +/- 1	3 +/- 1	3 +/- 1	6.60E-25
Maximum post bronchodilator FEV1 (% predicted)	105 +/- 11	99 +/- 14	104 +/- 12	86 +/- 13	82 +/- 14	70 +/- 20	5.30E-35
Prebronchodilator FEV1/FVC (absolute ratio)	1 +/- 0	1 +/- 0	1 +/- 0	1 +/- 0	1 +/- 0	1 +/- 0	1.30E-38
Use oral CSs (%)	100/0	99/1	84/16	85/15	40/60	4/96	6.00E-41
Use combination therapy (%)	100/0	59/41	16/84	29/71	23/77	25/75	8.40E-26
Sputum frequency (score)	1 +/- 1	2 +/- 1	3 +/- 1	3 +/- 2	3 +/- 2	4 +/- 2	4.30E-23
Symptoms worsen when CS decreased (%)	100/0	70/30	32/68	35/65	30/70	12/88	1.90E-25
Allergy symptoms in summer (score)	1 +/- 1	2 +/- 1	3 +/- 1	3 +/- 1	2 +/- 1	3 +/- 1	6.90E-22
Asthma symptoms caused by animal exposure (%)	99/1	27/73	26/74	35/65	47/53	46/54	2.40E-25

Take systemic CS for 50% or more of last year (%)	100/0	100/0	95/5	94/6	53/47	17/83	2.20E-41
Maximum post bronchodilator FEV/FVC (absolute ratio)	1 +/- 0	1 +/- 0	1 +/- 0	1 +/- 0	1 +/- 0	1 +/- 0	1.00E-27
Have history of nasal polyps (%)	100/0	97/3	95/5	99/1	13/87	79/21	2.30E-46
Had surgery to remove nasal polyps (%)	99/1	99/1	100/0	99/1	17/83	96/4	9.40E-54
Had allergy symptoms without cold or flu (%)	71/29	7/93	11/89	10/90	17/83	21/79	2.30E-28
Use nebulized beta-agonist (%)	100/0	95/5	74/26	66/34	50/50	17/83	5.40E-26
Prebronchodilator FVC (% predicted)	102 +/- 11	97 +/- 14	97 +/- 16	85 +/- 13	79 +/- 17	72 +/- 17	1.80E-27
Ethnicity_NonHispanic (%)	5/95	1/99	100/0	0/100	3/97	4/96	1.80E-51
Age (age)	29 +/- 11	31 +/- 11	33 +/- 10	39 +/- 12	47 +/- 10	43 +/- 11	2.90E-17
Number of skin reactions to allergens	1 +/- 2	4 +/- 3	4 +/- 3	4 +/- 3	2 +/- 2	4 +/- 4	1.60E-22
Visited ER for breathing in last 12 mos (%)	100/0	93/7	84/16	65/35	67/33	21/79	3.30E-22
Use leukotriene modifier (%)	100/0	91/9	63/37	65/35	57/43	25/75	6.50E-20
Atopy (%)	62/38	7/93	5/95	9/91	20/80	17/83	8.50E-22
Total WBC (x10^3 cells)	6 +/- 1	6 +/- 2	6 +/- 2	7 +/- 2	8 +/- 3	10 +/- 3	2.10E-17
Maximal bronchodilator reversibility (percent change)	6 +/- 3	12 +/- 9	16 +/- 23	19 +/- 17	17 +/- 15	36 +/- 37	1.40E-11
Number of ICU admissions for asthma (NULL)	0 +/- 0	0 +/- 0	0 +/- 1	0 +/- 1	0 +/- 1	1 +/- 1	6.70E-15
Exhaled NO (parts per billion)	20 +/- 16	34 +/- 24	29 +/- 29	39 +/- 32	40 +/- 26	82 +/- 64	1.30E-13
Hospitalized for asthma in last 12 mos (%)	100/0	99/1	89/11	92/8	77/23	38/62	4.60E-24
Ever had acute or recurrent sinusitis (%)	95/5	72/28	58/42	66/34	20/80	33/67	6.50E-15
Ever had recurrent bronchitis (%)	98/2	84/16	74/26	56/44	53/47	38/62	2.90E-14
Ever had sinus surgery (%)	99/1	93/7	84/16	91/9	30/70	83/17	4.60E-22
Use nasal steroids (%)	93/7	79/21	32/68	58/42	40/60	38/62	9.10E-14
BMI (kg/m^2)	26 +/- 7	27 +/- 5	29 +/- 9	31 +/- 7	31 +/- 6	30 +/- 7	1.80E-07
Use inhaled CSs (%)	100/0	84/16	84/16	78/22	67/33	33/67	2.90E-12
Blood neutrophil percentage (%)	55 +/- 9	57 +/- 9	59 +/- 10	56 +/- 9	63 +/- 14	70 +/- 15	5.60E-07
Ethnicity_Hispanic (%)	96/4	99/1	37/63	100/0	97/3	100/0	1.00E-28
Blood neutrophil numbers (x10^3)	3 +/- 3	4 +/- 2	4 +/- 2	4 +/- 1	5 +/- 3	7 +/- 3	6.90E-13
Blood lymphocyte percentage (%)	33 +/- 8	32 +/- 8	31 +/- 9	32 +/- 8	26 +/- 11	21 +/- 11	4.20E-14
Number of children (NULL)	1 +/- 1	0 +/- 1	1 +/- 1	1 +/- 1	2 +/- 1	2 +/- 2	1.60E-11
Maximum post bronchodilator FVC (% predicted)	104 +/- 12	103 +/- 13	103 +/- 15	95 +/- 13	93 +/- 13	89 +/- 20	9.60E-10
Systolic BP (mmHg)	120 +/- 14	122 +/- 14	114 +/- 11	128 +/- 16	134 +/- 15	128 +/- 18	2.00E-07
Height (cm)	171 +/- 10	172 +/- 9	162 +/- 9	168 +/- 9	172 +/- 9	166 +/- 10	6.60E-06
Ever had pneumonia (%)	87/13	76/24	74/26	49/51	77/23	21/79	4.00E-11
Use theophylline (%)	100/0	100/0	84/16	97/3	93/7	58/42	1.80E-18

Have GERD (%)	96/4	92/8	89/11	67/33	67/33	50/50	1.30E-10
Diagnosed with osteoporosis (%)	98/2	100/0	100/0	96/4	90/10	54/46	1.30E-18
Diastolic BP (mmHg)	72 +/- 9	73 +/- 9	74 +/- 7	77 +/- 11	80 +/- 8	79 +/- 14	1.50E-04
Blood eosinophil percentage (%)	3 +/- 2	4 +/- 3	4 +/- 3	5 +/- 4	4 +/- 4	3 +/- 4	4.00E-05
Blood eosinophil numbers (x10^3)	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	8.40E-07
Mother/Father with asthma (%)	93/7	69/31	63/37	51/49	87/13	62/38	3.90E-08
Blood lymphocyte numbers (x10^3)	2 +/- 0	2 +/- 1	2 +/- 1	2 +/- 1	2 +/- 1	2 +/- 1	9.50E-03
Use ipratropium (%)	100/0	99/1	89/11	95/5	97/3	58/42	4.40E-16
BAL eosinophil percentage (%)	0 +/- 0	1 +/- 2	0 +/- 0	1 +/- 3	2 +/- 2	3 +/- 5	7.60E-09
Ethnicity_Unknown (%)	100/0	100/0	63/37	100/0	100/0	96/4	2.00E-23
BAL lymphocyte percentage (%)	8 +/- 6	6 +/- 5	5 +/- 5	7 +/- 7	8 +/- 7	11 +/- 9	N.S.
Use anti-IgE therapy (%)	100/0	100/0	100/0	97/3	87/13	71/29	1.60E-12
Number of siblings (NULL)	2 +/- 2	2 +/- 1	3 +/- 2	3 +/- 2	3 +/- 3	3 +/- 2	9.10E-04
Use injectable CSs (%)	100/0	100/0	95/5	100/0	97/3	71/29	5.90E-16
Exposed to 2nd hand smoke (%)	89/11	83/17	79/21	52/48	87/13	71/29	1.80E-07
BAL alveolar macrophage percentage (%)	88 +/- 8	90 +/- 8	93 +/- 6	88 +/- 10	84 +/- 14	80 +/- 16	1.40E-05
Asthma worsened by nonsteroidal pain medications (%)	99/1	97/3	89/11	94/6	67/33	79/21	1.00E-08
Race_AfricanAmerican (%)	91/9	81/19	89/11	59/41	93/7	67/33	3.40E-06
Mother with asthma (%)	96/4	80/20	89/11	65/35	93/7	79/21	1.00E-05
Race_Caucasian (%)	21/79	23/77	58/42	47/53	7/93	33/67	5.70E-06
Blood monocyte numbers (x10^3)	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	1 +/- 0	N.S.
Diagnosed with chronic bronchitis (%)	100/0	99/1	95/5	86/14	90/10	79/21	1.60E-05
BAL neutrophil percentage (%)	3 +/- 3	3 +/- 4	2 +/- 2	3 +/- 5	6 +/- 9	7 +/- 9	N.S.
Use long acting beta-agonist (%)	100/0	96/4	89/11	90/10	83/17	71/29	1.60E-05
Blood monocyte percentage (%)	8 +/- 4	7 +/- 2	6 +/- 2	7 +/- 2	6 +/- 2	6 +/- 3	1.60E-05
Ever intubated (%)	100/0	98/2	89/11	94/6	90/10	71/29	7.40E-07
Blood basophil numbers (x10^3)	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	5.40E-03
Sibling with asthma (%)	88/12	73/27	63/37	56/44	73/27	67/33	4.00E-04
Had prolonged skin rash (%)	92/8	70/30	79/21	75/25	63/37	62/38	1.10E-03
Father with asthma (%)	97/3	87/13	68/32	78/22	93/7	79/21	1.20E-03
Hematocrit (%)	41 +/- 7	41 +/- 4	40 +/- 4	41 +/- 4	45 +/- 10	41 +/- 4	3.00E-02
Gender_Male (%)	59/41	57/43	84/16	75/25	40/60	75/25	2.30E-03
Gender_Female (%)	41/59	43/57	16/84	25/75	60/40	25/75	2.30E-03
Race_Multiracial (%)	95/5	99/1	79/21	95/5	100/0	100/0	5.00E-04
Diagnosed with diabetes (%)	100/0	99/1	95/5	97/3	93/7	83/17	7.60E-04
Father with allergy (%)	82/18	64/36	68/32	70/30	87/13	83/17	1.30E-02
Use alternative therapy (%)	100/0	99/1	100/0	92/8	90/10	96/4	1.40E-02
Ever smoke (%)	89/11	84/16	84/16	68/32	70/30	75/25	1.20E-02
Race_Other (%)	99/1	99/1	84/16	100/0	100/0	100/0	8.20E-05
Use immunotherapy (%)	100/0	99/1	100/0	100/0	90/10	100/0	1.60E-04

Use combination albuterol-ipratropium (%)	100/0	100/0	100/0	97/3	100/0	92/8	6.00E-03
Have a cat (%)	70/30	87/13	84/16	77/23	73/27	88/12	N.S.
Mother with allergy (%)	75/25	58/42	68/32	58/42	63/37	75/25	N.S.
Mother with COPD (%)	98/2	97/3	100/0	92/8	90/10	88/12	N.S.
Race_Asian (%)	95/5	99/1	89/11	99/1	100/0	100/0	N.S.
Smoking status (%)	90/10	84/16	79/21	76/24	73/27	75/25	N.S.
IgE level (IU/ml)	40 +/- 56	289 +/- 569	271 +/- 406	388 +/- 1362	173 +/- 216	253 +/- 466	3.60E-12
Total cell count in BAL (x10 ⁶)	13 +/- 45	9 +/- 11	9 +/- 6	10 +/- 8	7 +/- 5	3 +/- 3	6.30E-06
Have a dog (%)	77/23	83/17	68/32	75/25	80/20	67/33	N.S.
Have a bird (%)	98/2	98/2	95/5	97/3	93/7	100/0	N.S.
Ethnicity_Refused (%)	99/1	100/0	100/0	100/0	100/0	100/0	N.S.
Father with eczema (%)	96/4	93/7	95/5	96/4	97/3	96/4	N.S.
Mother with eczema (%)	93/7	92/8	89/11	92/8	93/7	88/12	N.S.

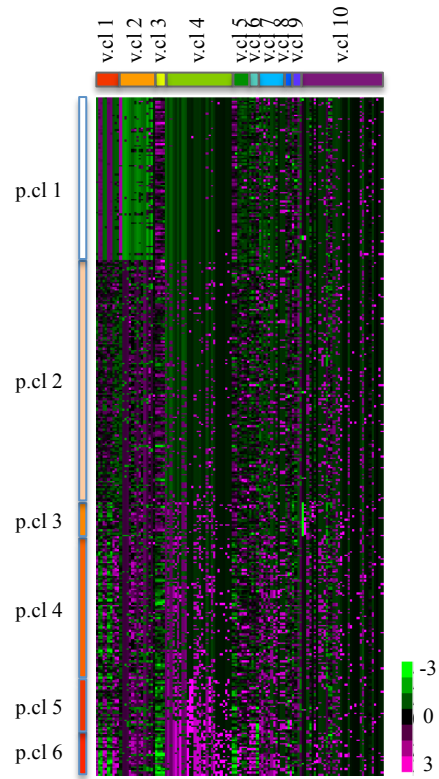


Figure E1. Heatmap of the 378 SARP subjects clustered into 6 groups (rows) and the 112 variables clustered into 10 categories (columns). The variables in each cluster are arranged in the same order, with decreasing INFOGAINs, as shown in Table E2.

Online Repository figure legends

Fig E1. Heatmap of the 378 SARP subjects clustered into 6 groups (*rows*) and the 112 variables clustered into 10 categories (*columns*). The variables in each cluster are arranged in the same order, with decreasing INFOGAINs, as shown in Table E2.

Fig E2. The dendrogram of the hierarchical tree of the variables in the SARP data. A hierarchical clustering algorithm with the Ward's criteria was employed to partition the variables into 10 clusters. The *color bars* represent the same variable clusters as those illustrated in Figure E1. The *red dashed line* shows the cutoff distance value used for determining the number of the clusters. Note that the orderings of the leaf nodes (i.e., variables) and clusters in this plot were generated by the hierarchical clustering algorithm, which is thus different from the orderings of the variables and clusters in Tables E2 and E3, and Figure E1 ordered by INFOGAINs. The variables (represented as the leaf nodes here) correspond to the indexes of the variables shown in Table E3 in the following order: 16, 30, 9, 3, 4, 8, 10, 5, 6, 2, 33, 13, 1, 20, 11, 7, 12, 38, 17, 32, 25, 18, 114, 46, 43, 54, 36, 37, 75, 67, 84, 93, 49, 51, 22, 44, 79, 69, 23, 34, 28, 39, 24, 31, 29, 21, 19, 14, 91, 57, 52, 45, 55, 15, 26, 40, 63, 48, 27, 35, 71, 72, 76, 50, 98, 65, 99, 100, 47, 60, 58, 61, 105, 113, 56, 64, 70, 66, 53, 89, 68, 42, 62, 90, 78, 83, 86, 73, 110, 120, 102, 81, 108, 115, 97, 103, 111, 96, 119, 101, 59, 41, 87, 85, 92, 88, 104, 107, 109, 80, 95, 77, 112, 116, 106, 117, 118, 82, 74, 94.

Fig E3. Pre and post bronchodilator lung function variables differed by subject clusters. **A-C**, Cluster 6 had lower prebronchodilator FEV1% predicted (**A**), prebronchodilator FVC% predicted (except Cluster 5; **B**), and Maximum post bronchodilator FEV1% predicted (**C**) than all the other clusters. Clusters 4 and 5 were lower than clusters 1-3, but clusters 4 and 5 did not differ from each other. Cluster 3 did not differ from cluster 1. **D**, Clusters 4-6 had higher maximum post bronchodilator FEV1% predicted than clusters 1-2. Cluster 3 did not differ from clusters 1 and 2.

Fig E4. Non-eosinophilic blood inflammatory markers. **A-B**, Cluster 6 had higher total WBC (**A**) and blood neutrophil numbers (**B**) compared to all the other groups, while cluster 3 did not differ from cluster 1. Cluster 5 was also higher than clusters 1-3 (**A**) and higher than clusters 1, 2 and 4 (**B**).

Fig E5. BAL inflammatory cell types. **A**, Cluster 5 had higher BAL eosinophils differing from all other clusters except cluster 6. BAL eosinophils were lower in cluster 1 than all other groups, but clusters 2-4, and 6 did not differ from one another. **B**, Cluster 6 had higher BAL neutrophil percentages than clusters 2 and 3, but the differences were not significant ($P = .09$). **C**, Cluster 6 had higher lymphocyte percentages than clusters 2-5, but the differences were not significant.

Fig E6. Ethnicity and heredity. **A**, Cluster 3 had highest Hispanic and lowest non-Hispanic subjects among all clusters. **B**, Cluster 4 had a higher proportion of subjects with a family history of asthma, being higher than clusters 1, 2 and 5. Cluster 5 did not differ from cluster 1.

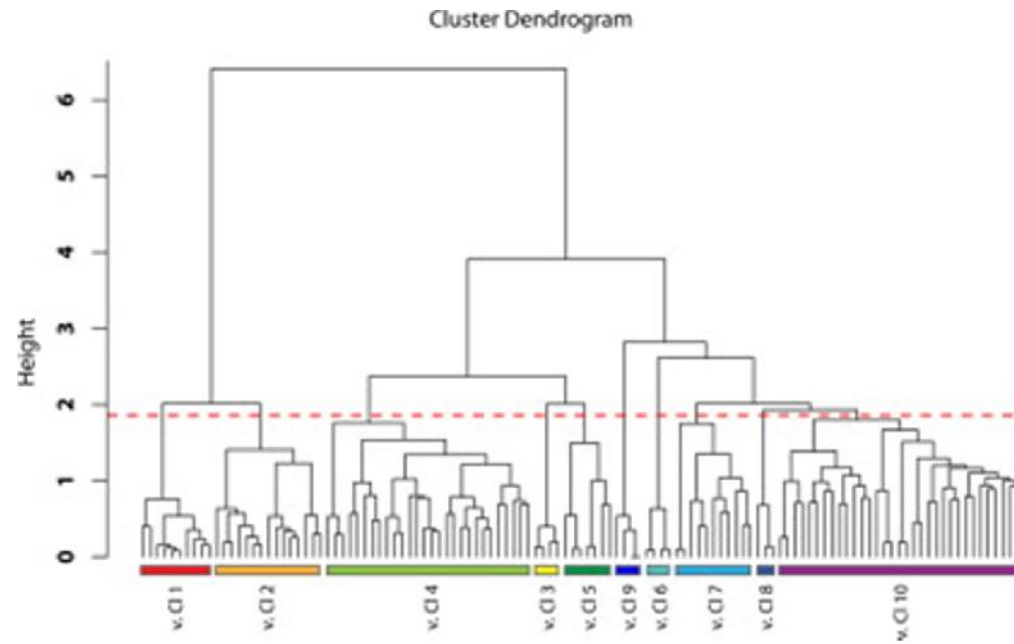


Fig E2

Wei Wu , Eugene Bleeker , Wendy Moore , William W. Busse , Mario Castro , Kian Fan Chung , William J. Calhoun , ...

Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data

Journal of Allergy and Clinical Immunology, 2014

<http://dx.doi.org/10.1016/j.jaci.2013.11.042>

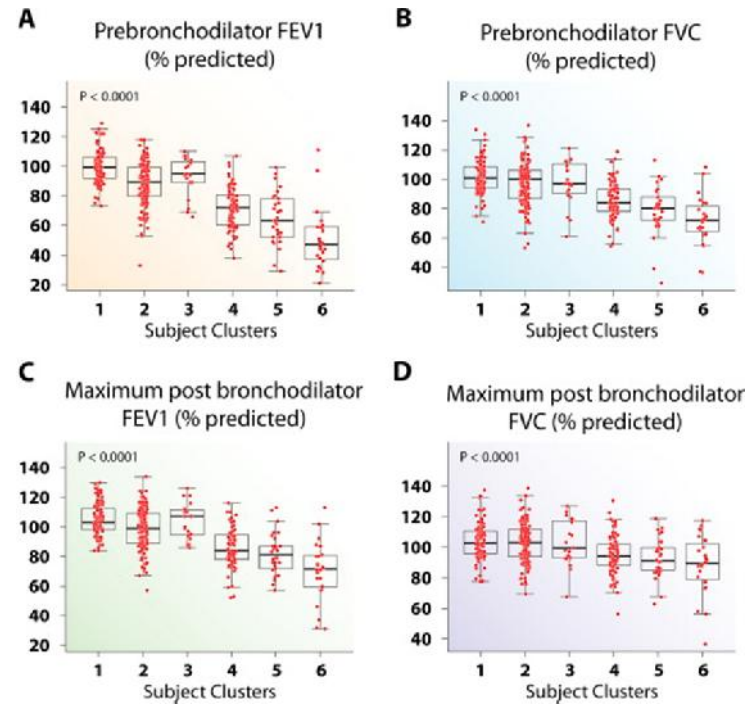


Fig E3

Wei Wu , Eugene Bleeker , Wendy Moore , William W. Busse , Mario Castro , Kian Fan Chung , William J. Calhoun , ...

Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data

Journal of Allergy and Clinical Immunology, 2014

<http://dx.doi.org/10.1016/j.jaci.2013.11.042>

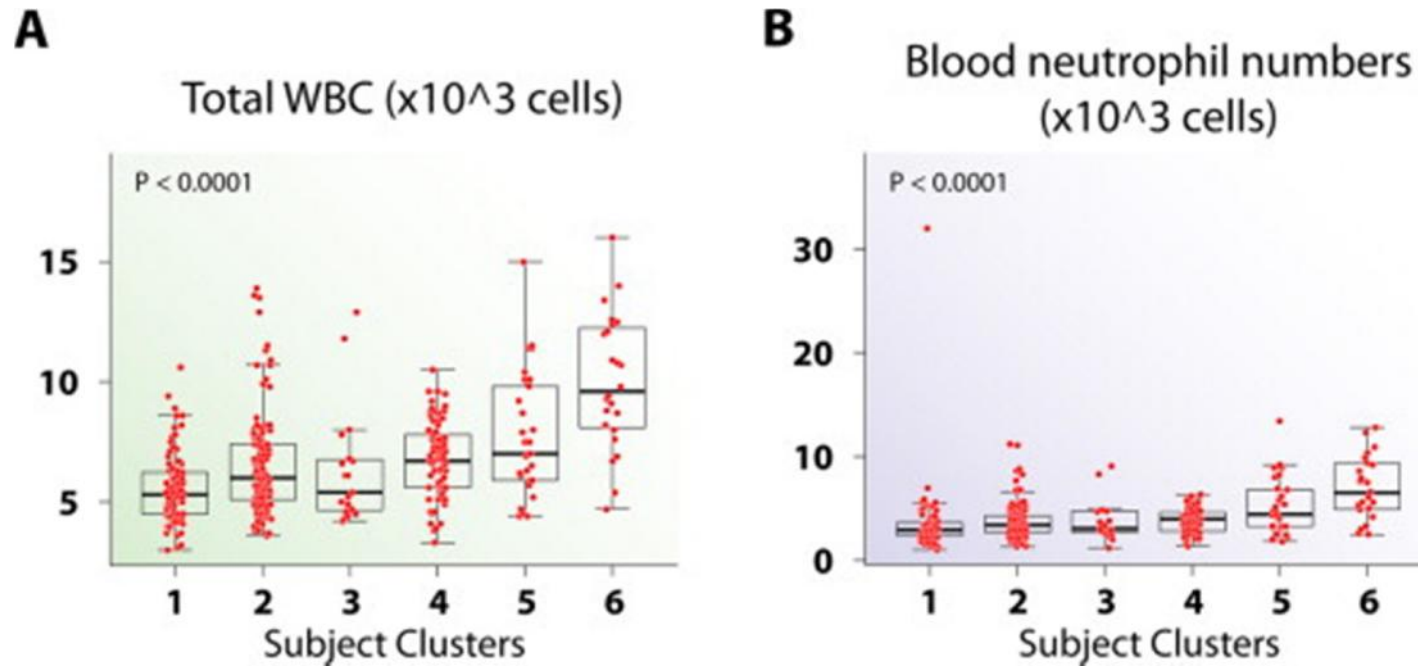


Fig E4

Wei Wu , Eugene Bleeker , Wendy Moore , William W. Busse , Mario Castro , Kian Fan Chung , William J. Calhoun , ...

Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data

Journal of Allergy and Clinical Immunology, 2014

<http://dx.doi.org/10.1016/j.jaci.2013.11.042>

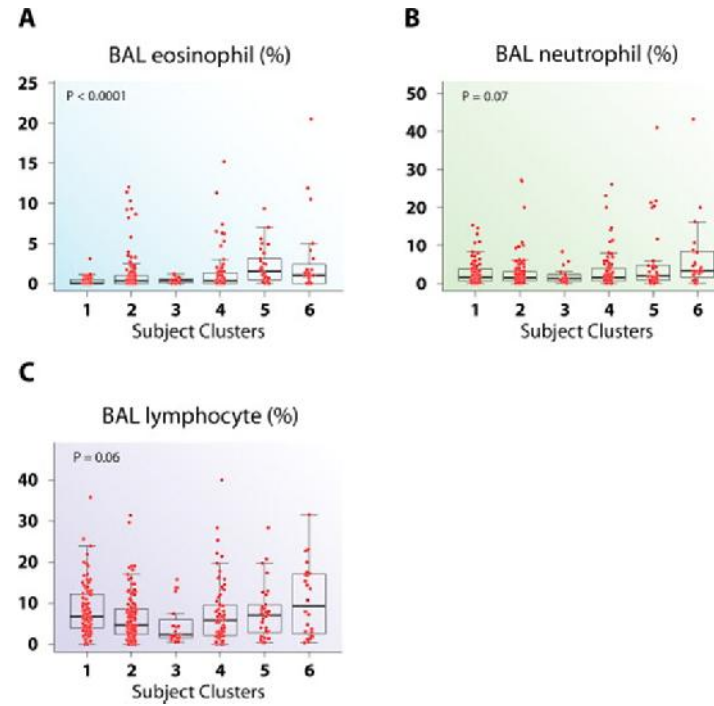


Fig E5

Wei Wu , Eugene Bleeker , Wendy Moore , William W. Busse , Mario Castro , Kian Fan Chung , William J. Calhoun , ...

Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data

Journal of Allergy and Clinical Immunology, 2014

<http://dx.doi.org/10.1016/j.jaci.2013.11.042>

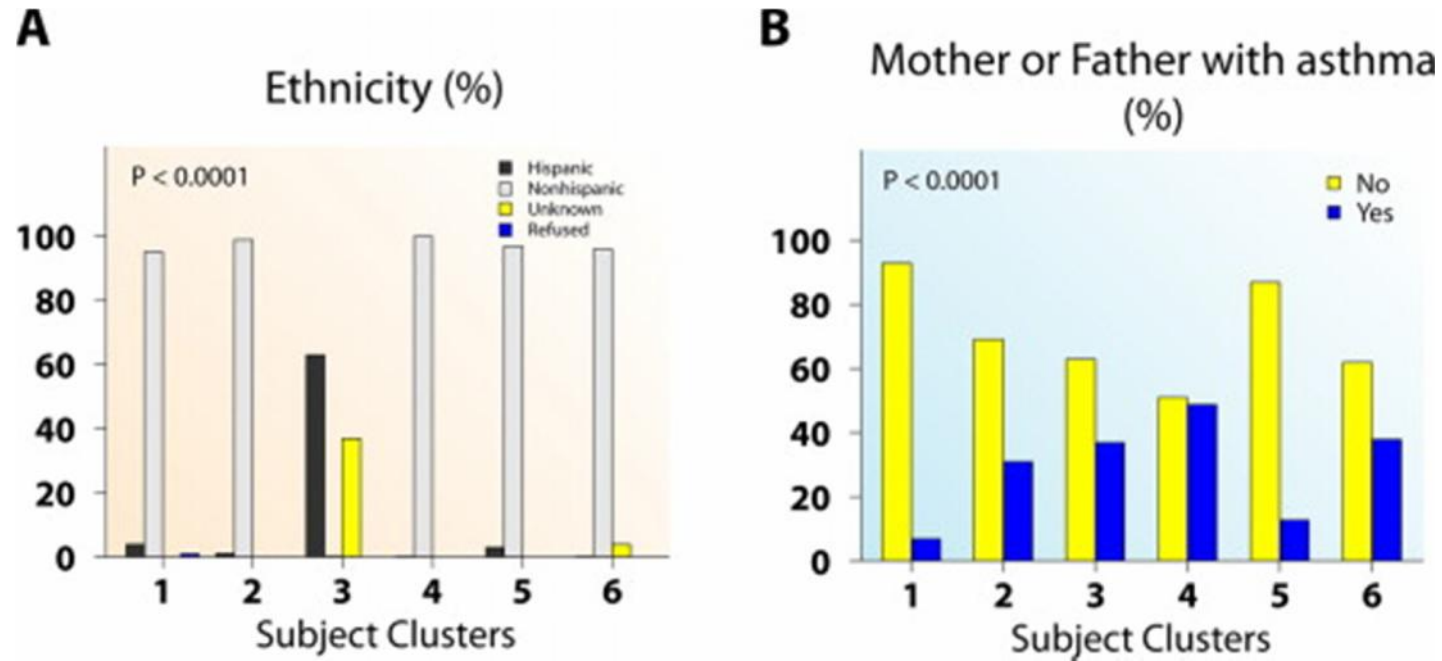


Fig E6

Wei Wu , Eugene Bleecker , Wendy Moore , William W. Busse , Mario Castro , Kian Fan Chung , William J. Calhoun , ...

Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data

Journal of Allergy and Clinical Immunology, 2014

<http://dx.doi.org/10.1016/j.jaci.2013.11.042>