Supplementary material

Identifying adult asthma phenotypes using a clustering approach

Valérie Siroux[1,2], Xavier Basagaña[3,4,5,6], Anne Boudier[1,2], Isabelle Pin[1,2,7], Judith Garcia-Aymerich[3,4,5,6], Aurélien Vesin[1,2], Rémy Slama[1,2], Deborah Jarvis[8], Josep M Anto[3,4,5,6], Francine Kauffmann[9,10], Jordi Sunyer[3,4,5,6]

1. Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Inserm, U823, Grenoble, France
2. Université Joseph Fourier, Grenoble, France
3. Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
4. Institut Municipal d'Investigació Mèdica (IMIM-Hospital del Mar), Barcelona, Spain
5. Department of Experimental Sciences and Health, Universitat Pompeu Fabra, Barcelona, Spain
6. CIBER en Epidemiologia y Salud Pública (CIBERESP), Barcelona, Spain
7. CHU Grenoble, pédiatrie, Grenoble, France
8. Respiratory Epidemiology and Public Health Group, Imperial College, London, United-Kingdom
9. Respiratory and Environmental epidemiology, CESP Centre for research in Epidemiology and Population health, U1018, Inserm, F-94807, Villejuif, France
10. Université Paris Sud 11, UMRS 1018, F-94807,Villejuif, France

Methods

The ECRHS is an international study of asthma (1991-1993) [1]. From a short questionnaire sent to a random sample of young adults aged 20-44 years in 48 centres, a random sample of the responders enriched with a symptomatic sample was examined in ECRHS I (n = 18356) (see Figure E1 in the online data supplement). About 10 years later, 10933 individuals in 29 centres from the initial cohort were re-examined (ECRHS II) [2]. The EGEA study combines a case-control and family study of adults and children with asthma. The first EGEA survey (EGEA1) included 388 cases with asthma, recruited in five chest clinics, their first-degree relatives and 415 population-based controls (n=2047) [3,4]. A twelve-year follow-up of this population was conducted from 2003 to 2007 (EGEA2, n=1601) [5] (see Figure E2 in the online data supplement). Written consent was obtained from all participants at both surveys, for the two studies. Ethical approval to carry out the study was obtained for both surveys from the relevant committees (Cochin Royal Hospital, Paris for EGEA1, Necker-Enfants Malades Hospital, Paris for EGEA2, from the Local ethics committees at each centre for ECRHS).

The precise definition of the nineteen variables firstly selected for the analysis is presented in Table E1.

In order to identify distinct asthma phenotypes Latent Class Analysis (LCA), a latent variable model that serves to cluster subjects into classes has been used [6]. This approach allows identifying a set of latent classes of individuals who are similar to each other based on the variables used in the analysis. Parameters estimated in LCA include class membership probabilities, which represent the proportion of a population expected to belong to each latent class, and conditional variable response probabilities, which represent the probability that an individual demonstrates a specific response to an observed variable, conditional on class membership. These parameters are combined mathematically into a likelihood function, and

the set that maximizes the likelihood function determines the membership of each individual in one of the latent classes. Compared to the other clustering approaches, the advantages of using LCA are 1) it is a well designed method for categorical input variables; 2) using a statistical model-based approach, the choice of a cluster criterion to assign individuals to cluster is less arbitrary and 3) all subjects are included in the analysis since missing data for the variables are handled in the procedure, with data assumed to be missing at random [7]. The test of the null hypothesis that data were missing at random (Missing Completely At Random test in the SAS LCA procedure) was non-significant in each dataset indicating that missingness does not affect the interpretation of the results [7].

The hypothesis when using the LCA is the « conditional independence » or « local independence »; that is that within each latent class, all input variables are statistically independent of each other. In order to avoid conditional dependence, an explanatory factor analysis (with varimax rotation) was first applied to 18 selected variables in order to identify variables representing the same dimension. Such variables would be likely to be dependent within all latent classes. Airway hyperresponsiveness (AHR) (PD20≤1mg metacholine) was not included in the factor analysis because it was missing for all subjects with low lung function (FEV1 < 70% for ECRHSII and <80% for EGEA2, precluded individuals from undergoing bronchial challenges). The factor analysis conducted in both the ECRHSII (table E2) and the EGEA2 (table E3) data identified an “asthma symptom” factor as the first factor explaining 18% of the variance in both studies. Consequently, five asthma symptoms (wheeze and breathless, woken up with a feeling of chest tightness, attack of shortness of breath at rest, attack of shortness of breath after exercise and woken by attack of shortness of breath) were combined in an asthma symptoms score as it has been previously published in ECRHS [8]. Based on the distribution of the score in the populations, the score was further divided in three

classes: 0 symptom, 1 or 2 symptoms, 3 or more symptoms. The second factor identified different dimensions in the two studies and had a much lower percentage of variance explained (<10% in EGEA2). Therefore only results from factor one were used to reduce the variable dimensionality. Finally, 15 variables were selected to be included in the LCA: age, sex, age of asthma onset, woken up by attack of coughing, asthma symptom score, chronic cough or phlegm, asthma attacks in the past 12 months, asthma exacerbation, the type of asthma treatment, eczema, rhinitis, atopy, total IgE,FEV1 and AHR. As atopy was defined on the basis of the response to 11 allergens in EGEA2 compared to only 4 allergens in ECRHSII, we checked that restricting the definition of atopy in EGEA2 on the 4 allergens tested in ECRHSII (cat, Dermatophagoides pteronyssinus, Cladosporium and timothy grass) only slightly changed the prevalence of atopy in EGEA2 (74.3% vs 79.7%) and did not change the latent classes observed (data not shown).

To determine the number of latent classes, models with different numbers of latent classes were compared using the Bayesian Information Criterion (BIC) and the model with the lowest BIC was selected. As the solution may depend on the starting value of the algorithm, we conducted the analysis with a large number of different random starting values and selected the model with the highest goodness-of-fit. Each subject was assigned to the latent class for which he had the highest membership probability [9]. The different latent classes were interpreted as phenotypes based on the model variables that characterize the best each latent class.

The descriptive and the latent class analyses were performed using the SAS 9.1 statistical software (SAS institute, Cary, NC). Summary statistics are reported as mean values and standard deviation (sd) for continuous variables and percentages for categorical variables. The SAS LCA procedure was used (available at http://methodology.psu.edu/).

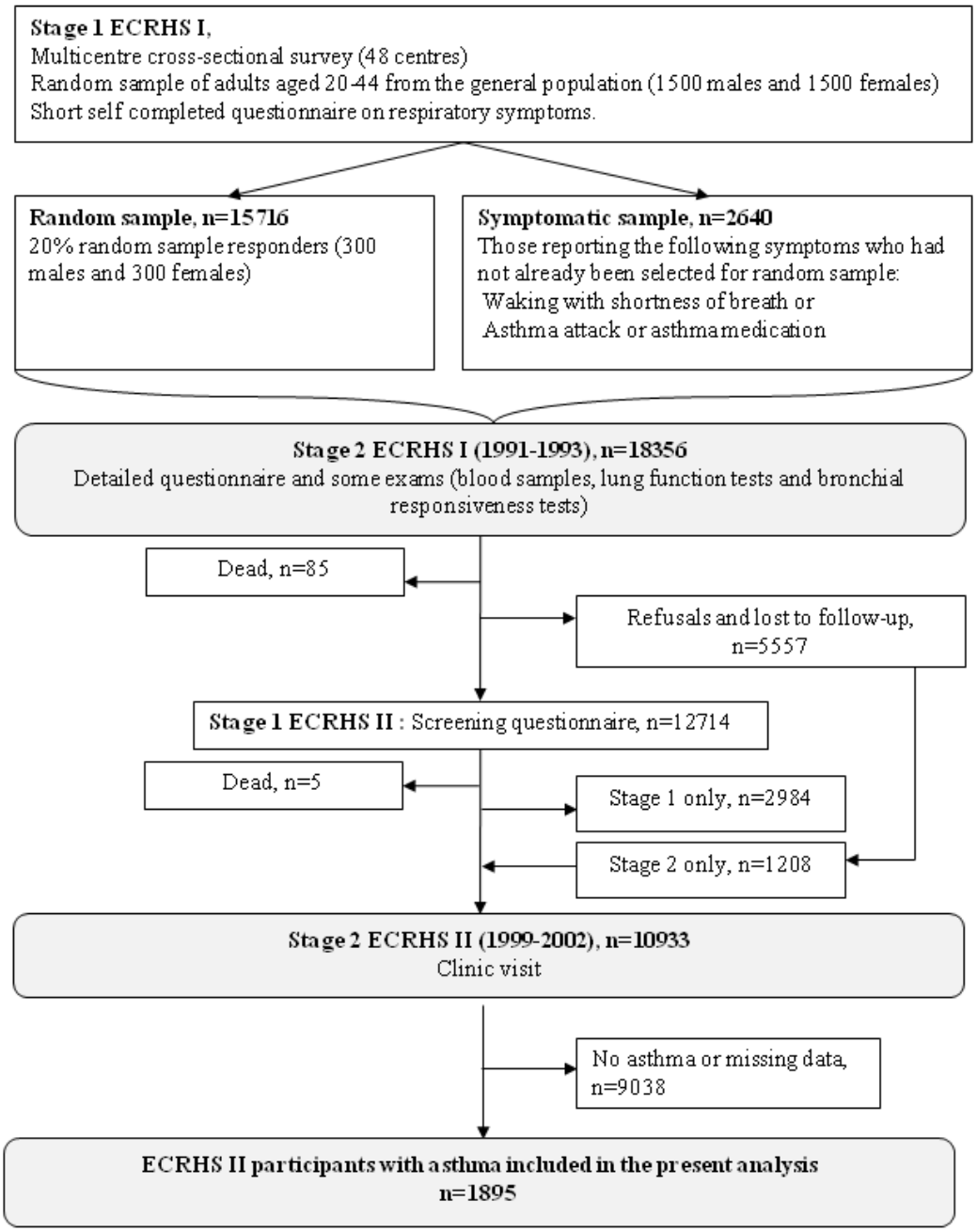

Figure E1: Flowchart of the ECRHS population

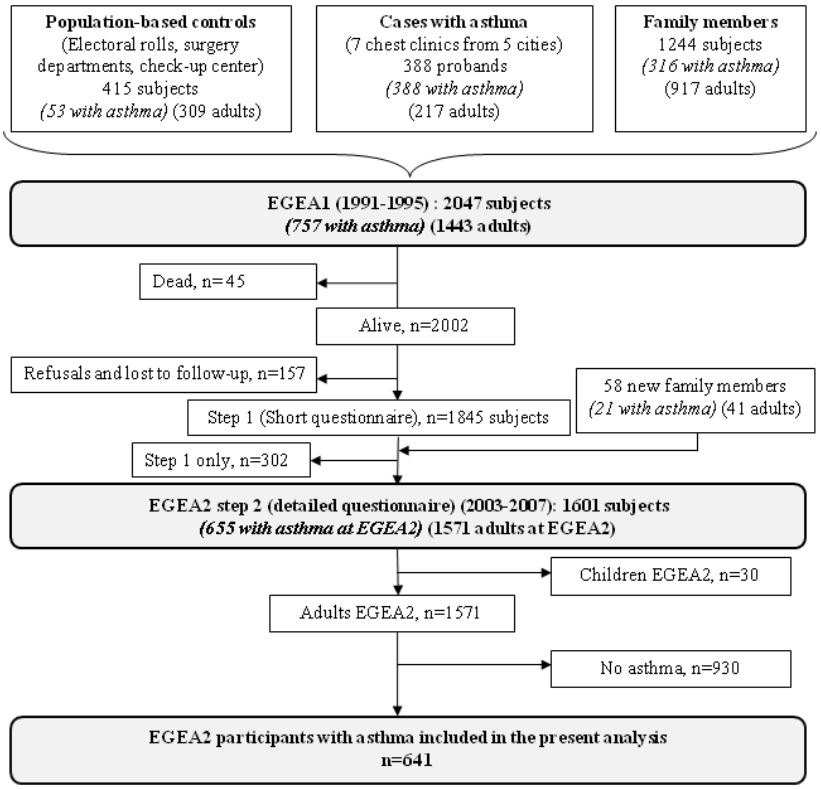

Figure E2: Flowchart of the EGEA population

REFERENCES

1. Burney PG, Luczynska C, Chinn S, Jarvis D. The European Community Respiratory Health Survey. *Eur Respir J* 1994;7:954-960.

2. The European Community Respiratory Health Survey II. *Eur Respir J* 2002;20:1071-1079.

3. Kauffmann F and Dizier MH. EGEA (Epidemiological study on the Genetics and Environment of Asthma, bronchial hyperresponsiveness and atopy)--design issues. EGEA Co-operative Group. *Clin Exp Allergy* 1995;25 Suppl 2:19-22.

4. Kauffmann F, Dizier MH, Pin I, Paty E, Gormand F, Vervloet D, Bousquet J, Neukirch F, Annesi I, Oryszczyn MP, Lathrop M, Demenais F, Lockhart A, Feingold J. Epidemiological study of the genetics and environment of asthma, bronchial hyperresponsiveness, and atopy: phenotype issues. *Am J Respir Crit Care Med* 1997;156:S123-S129.

5. Siroux V, Boudier A, Bousquet J, Bresson JL, Cracowski JL, Ferran J, Gormand F, Just J, Le Moual N, Morange S, Nadif R, Oryszczyn MP, Pison C, Scheinmann P, Varraso R, Vervloet D, Pin I, Kauffmann F. Phenotypic determinants of uncontrolled asthma. *J Allergy Clin Immunol* 2009;124:681-687.

6. Hagenaars J, Mc Cutcheon A. Applied latent class analysis. Cambridge University Press, 2002.

7. Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A sas procedure for latent class analysis. *Struct Equation Model* 2007;14:671-694.

8. Pekkanen J, Sunyer J, Anto JM, Burney P. Operational definitions of asthma in studies on its aetiology. *Eur Respir J* 2005;26:28-35.

9. Spycher BD, Silverman M, Brooke AM, Minder CE, Kuehni CE. Distinguishing phenotypes of childhood wheeze and cough using latent class analysis. *Eur Respir J* 2008;31:974-981.

TABLE E1: Definition of the variables firstly selected for the analysis.

Criteria	Definition used in ECRHS II	Definition used in EGEA2
Age	<40; ≥40	<40; ≥40
Age of asthma onset	"How old were you when you had your first attack of asthma" 3 classes: ≤4, 4-16, >16 years	idem
Wheezing with breathlessness, 12 months	"Have you had wheezing or whistling in your chest at any time in the last 12 months?" and "Have you been at all breathless when the wheezing the wheezing noise was present?"	idem
Woken up with feeling of tightness, 12 months	"Have you woken up with a feeling of tightness in your chest at any time in the last 12 months?"	idem
Attack of shortness of breath at rest, 12 months	"Have you had an attack of shortness of breath that came on during the day when you were at rest at any time in the last 12 months?"	idem
Shortness of breath during activity, 12 months	"Have you had an attack of shortness of breath that came on following strenuous activity at any time in the last 12 months?"	Idem
Woken up shortness of breath, 12 months	"Have you been woken by an attack of shortness of breath at any time in the last 12 months?"	Idem
Woken up by attack of coughing, 12 months	"Have you been woken by an attack of coughing at any time in the last 12 months?"	idem
Chronic cough or phlegm	"Do you usually cough during the day, or at night, in the winter, on most days for as much as three months each year" or "Do you usually bring up any phlegm from your chest during the day, or at night, in the winter, on most days for as much as three months each year?"	idem
Asthma attack, 12 months	"Have you had an attacks of asthma in the past 12 months"	idem

Exacerbation, 12 months	"Have you visited a hospital casualty department or emergency room because of breathing problems in the last 12 months?" or "Have you spent a night in hospital because of breathing problems in the last 12 months?" or "have you used oral corticosteroids to help your breathing at any time in the last 12 months?"	"Have you visited a hospital casualty department or emergency room because of asthma in the last 12 months?" or "Have you spent a night in hospital because of asthma in the last 12 months?" or "have you used oral corticosteroids to help your breathing at any time in the last 12 months?"
Asthma treatment, 3months	Assessed from a detailed questionnaire on the asthma treatment used in the past 3 months. 3 classes : no treatment, asthma treatment but not used daily, daily asthma treatment	Idem
Eczema	"Have you ever had eczema or any kind of skin allergy?"	"Have you ever had eczema?"
Rhinitis	"Do you have any nasal allergy including hay fever?"	"Have you ever had allergic rhinitis or hay fever?"
Atopy	Specific IgE to one allergen among cat, *Dermatophagoides pteronyssinus, Cladosporium* and timothy grass	Skin prick test to any of the 11 allergens: cat, *Dermatophagoides pteronyssinus, Blattela germanica,* olive, birch, *Parieteria judaica*, timothy grass, ragweed pollen, *Aspergillus, Cladosporium herbarum, Alternaria tenuis*)
Total IgE	<100IU/ml vs ≥100 IU/ml	idem
FEV1	Assessed using the best of five expiratory curves 2 classes : <80% predicted vs ≥80%	Idem
Airway hyperresponsiveness	For subjects with FEV1>70%, PD20≤1mg methacholine	For subjects with FEV1>80%, PD20≤1mg methacholine

TABLE E2 : Factor loadings of a factor analysis with varimax rotation in ECRHSII (n=1070)

	Factor1	**Factor2**
	18.3%	9.8%
Age (≥40yrs)	-0.18	0.20
Sex	0.04	0.17
Age of asthma onset (<4; 4-16; ≥16)	0.04	-0.10
Wheezing with breathlessness, 12 months	**0.64**	0.14
Woken up with feeling of tightness, 12 months	**0.70**	0.14
Attack shortness of breath at rest, 12months	**0.69**	0.03
Shortness of breath during activity, 12months	0.41	0.45
Woken up shortness of breath, 12months	**0.70**	0.16
Woken up by attack of coughing, 12months	0.25	**0.61**
Chronic cough or phlegm	0.13	**0.64**
Asthma attack, 12months	**0.71**	0.02
Exacerbation, 12 months	0.41	-0.16
Treatment (None; other than daily ICS; daily ICS)	0.22	0.12
Rhinitis	0.08	-0.08
Eczema	-0.16	0.28
Atopy	0.04	-0.42
Total IgE (≥100 IU/ml)	0.11	-0.17
FEV1 (<80 % predicted)	0.02	0.24

Only factors with %variance explained ≥10% were selected

TABLE E3. Factorial analysis with varimax rotation in EGEA2 (n=447)

	Factor1	**Factor2**
	18.8%	12.4%
Age (≥40yrs)	-0.04	**0.78**
Sex	0.23	-0.13
Age of asthma onset (<4; 4-16; ≥16)	-0.15	0.53
Wheezing with breathlessness, 12 months	**0.60**	-0.09
Woken up with feeling of tightness, 12 months	**0.70**	-0.02
Attack shortness of breath at rest, 12months	**0.68**	0.05
Shortness of breath during activity, 12months	0.35	-0.09
Woken up shortness of breath, 12months	**0.69**	-0.06
Woken up by attack of coughing, 12months	0.26	-0.09
Chronic cough or phlegm	0.05	0.24
Asthma attack, 12months	**0.70**	0.07
Exacerbation, 12 months	0.54	0.19
Treatment (None; other than daily ICS; daily ICS)	0.48	0.50
Rhinitis	0.09	-0.08
Eczema	0.01	-0.08
Atopy	0.13	-0.15
Total IgE (≥100 IU/ml)	0.12	-0.10
FEV1 (<80 % predicted)	0.17	**0.69**

Only factors with %variance explained ≥10% were selected