ERJ Express. Published on January 23, 2008 as doi: 10.1183/09031936.00153507

Distinguishing phenotypes of childhood wheeze and cough using latent class analysis

Full names, institution and country of all co-authors
Ben Daniel Spycher[1]
Michael Silverman[2]
Adrian Mark Brooke[2]
Christoph Erwin Minder[1]
Claudia Elisabeth Kuehni[1]
1: Swiss Pediatric Respiratory Research Group, Department of Social and Preventive Medicine, University of Bern, CH - 3012 Bern, Switzerland
2: The Leicester Children's Asthma Centre, Division of Child Health, Department of Infection, Immunity & Inflammation, University of Leicester, Leicester, LE2 7LX, UK

Address for correspondence

Dr. Claudia E. Kuehni, Swiss Paediatric Respiratory Research Group, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, CH-3012, Bern, Switzerland.
Phone: +41 (0)31 631 35 07, Fax: +41 (0)31 631 35 20, e-mail: kuehni@ispm.unibe.ch

Funding:
The work presented in this paper was funded by the Swiss National Science Foundation (PROSPER grant 3233-069348 and 3200-069349, and SNF grant 823B - 046481) and the Swiss Society of Pneumology. Original data collection was funded by the UK National Asthma Campaign. Follow-up data collection was funded by grants from: University Hospitals of Leicester NHS Trust (R&D), Leicestershire & Rutland Partnership Trust, Medisearch, Trent NHS Regional Health Authority, and the UK Department of Health (grant 0020014).
Running title: Identifying phenotypes of childhood asthma
Word count: 3,623
Keywords: Airway function, allergy, wheeze/asthma, bronchial responsiveness, cluster analysis, latent class modelling.

Abstract

Airway disease in childhood comprises a heterogeneous group of disorders. Attempts to distinguish different phenotypes have generally considered few disease dimensions. This study examines phenotypes of childhood wheeze and chronic cough, by fitting a statistical model to data representing multiple disease dimensions.

From a population-based, longitudinal cohort study of 1650 preschool children, 319 with parent-reported wheeze or chronic cough were included. Phenotypes were identified by latent class analysis using data on symptoms, skin-prick tests, lung function and airway responsiveness from two preschool surveys. These phenotypes were then compared with respect to outcome at school age.

The model distinguished three phenotypes of wheeze and two phenotypes of chronic cough. Subsequent wheeze, chronic cough and inhaler use at school age differed clearly between the five phenotypes. The wheeze phenotypes shared features with previously described entities and partly reconciled discrepancies between existing sets of phenotype labels.

This novel multidimensional approach has the potential to identify clinically relevant phenotypes not only in paediatric disorders but also in adult obstructive airway diseases, where phenotype definition is an equally important issue.

Introduction

It is widely accepted that childhood asthma comprises several distinct disorders, characterized by the common symptom of wheeze [1-4]. Distinguishing between these disorders is clinically important since aetiology, pathophysiology, potential for therapy and outcome may differ [1, 5-7]. Similarly, it has been emphasised that, although some children with chronic cough might suffer from a variant form of asthma, "lumping" together all chronic coughers under the term "cough variant asthma" is probably wrong [8].

Obstructive airway diseases clearly have multiple dimensions which involve atopy, disordered lung function, airway responsiveness and a variety of symptoms. Despite this, traditional phenotype definitions have used simple distinctions, such as a clinical classification into "exclusive viral wheeze" triggered only by colds and "multiple trigger wheeze" triggered also by other factors [9], or a retrospective classification by symptom history into "early transient", "persistent" and "late-onset" wheeze [2, 3, 7]. Because they are limited to single dimensions, such phenotype definitions embody an arbitrary element and may not properly reflect underlying disease processes. Furthermore, it is unclear how the different sets of phenotype labels relate to each other and whether they identify similar entities. For instance, is "exclusive viral wheeze" the same condition as "early transient wheeze"? We still lack an agreed system of classification that appropriately reflects underlying disease processes and, potentially, therapeutic responses. It has been proposed that statistical methods which can account for multiple dimensions of airway disease may facilitate the identification of relevant phenotypes [10].

Latent class analysis (LCA) [11, 12] is a statistical method developed in the social sciences which is used to identify distinct subsets (classes) of a population. The underlying classes are not observable and must be determined from the observed data. LCA has recently been used in medical research to identify disease phenotypes [13, 14]. The aims of the present study were (i) to apply LCA to a multivariate dataset combining symptoms and physiological measurements in order to identify and describe phenotypes of wheeze and cough in childhood, and (ii) to explore the validity of the resultant phenotypes by assessing how well they predicted future outcomes. The emphasis of the present paper is on the potential of this approach to identify phenotypes of obstructive airway disease.

Materials and Methods

Subjects and study design

In a population-based cohort study of 1650 white children recruited in 1990 at the age of 0 to 5 years in Leicestershire, UK [15-20], parents completed postal questionnaires on respiratory symptoms, exposures and socio-demographic characteristics in 1990, 1998 and 2003. Between 1992 and 94, a nested sample of 795 children was invited for physiological measurements and interviews [17, 18], including all with parent-reported wheeze (n=222) or chronic cough (cough occurring apart from colds n=226) in 1990 and a random sample of previously asymptomatic children (n=347). The study was approved by the Leicester Health Authority Committee on the Ethics of Clinical Research Investigation.

Identification of phenotypes was based on data from the first two surveys (1990 and 1992-4). From among the 488 respondents to the second survey (1992-4) we analyzed data from all those with a positive response in either survey (1990 or 1992-4) to one or both of the questions: "Has your child ever had attacks of wheezing?" and "Does he/she usually have a cough apart from colds?" (n=319) (Figure 1).

In a next step we compared prognosis between identified phenotypes, using data on current (i.e. previous 12-month) wheeze, frequent wheeze, bronchodilator use and cough without colds from two recent surveys, 1998 and 2003, when the children were aged 8-13 and 13-18 years respectively. Children who were asymptomatic in the first two surveys (n=169) served as a control group.

Physiological measurements

Physiological measurements included in this analysis were age- and height-standardized z-scores [21] of the pre-bronchodilator forced expiratory volume in 0.5s ($FEV_{0.5}$), bronchial responsiveness (provoking concentration of methacholine causing a 20% decrease in transcutaneous oxygen tension ($PC20tc\text{-}PO_2$)) [22], and atopy assessed by skin prick testing. Subjects responding to one or more of four aeroallergens (cat hair, dog danders, *Dermatophagoides pteronyssinus* and mixed grass pollen) were designated atopic. For more details see online supplementary material.

Statistical analysis

To identify phenotypes, LCA was applied to a set of variables measured on the sample of 319 children during the first two surveys. LCA assumes that the population is composed of subpopulations (latent classes), each having its distinctive distribution of the included variables [11]. If these variables represent disease manifestations the latent classes can be interpreted as clinical phenotypes. Application of LCA involves some prior decisions: (a) choosing the variables and (b) the number of latent classes to be included in the model. When choosing which variables to include there has to be a balance between using all potentially relevant information and the need to limit the number of parameters in the model. In the present study all parent-reported symptom data relating to cough and wheeze from the first two surveys and all measurements of atopy, lung function and bronchial responsiveness were considered for inclusion. Multiple correspondence analysis [23] was then used to make a narrower selection. In addition we included the variables age and sex (for a list of all included variables see tables 1 and 2). In order to choose the appropriate number of latent classes the model was repeatedly fitted with the number of classes increasing stepwise from 1 (model 1) to 7 (model 7). These models were then compared using bootstrapped p-values for the likelihood ratio test and the Bayesian information criterion [11].

The model was fitted by maximum likelihood estimation using Multimix, a Fortran program designed to fit latent class models including both continuous and categorical variables [24]. The variables $FEV_{0.5}$ and log transformed $tc\text{-}PO_2$ [25] were treated as continuous with a normal distribution and all other variables as categorical. We adapted the program to deal with missing data [26] and conditional questions (such as questions on shortness of breath, or seasonality of symptoms which were asked only to those children reporting wheeze ever). For more details on the modelling approach see the online supplementary material.

LCA allows computing the probability of belonging to a particular phenotype given the observed features of a subject. As is common practice in LCA [11], each child in the sample was assigned to the phenotype for which it had the highest membership probability. We refer to groups of children assigned in this way to different phenotypes as “phenotype clusters”. Two-sided Fisher’s exact tests were used to test associations between phenotype clusters and prognostic endpoints. These were computed using Stata statistical software (version 8.2, STATA Corporation, College Station, TX). A Bonferroni-corrected significance level was used to account for multiple pair-wise testing.

Results

Sample characteristics

The sample used for phenotype definition (n = 319) consisted of 189 (59%) children with wheeze ever reported in 1990 and/or in 1992-4 and 130 (41%) children with cough apart from colds reported in at least one survey, but no wheeze ever. The sample contained 160 (50%) girls and the median age (range) was 3.3 (0.3-5.4) years in 1990 and 6.3 (4.1 to 8.8) years in 1992-4. The healthy comparison group consisted of 169 asymptomatic children.

Phenotype identification

The two criteria which were applied to determine the number of phenotypes did not agree: the bootstrapped p-values for the LR test indicated five phenotypes (model 5) while the BIC preferred only two (model 2). Because this method is explorative and has the potential to reveal new phenotypes we chose to present model 5 (tables 1 and 2), knowing that the heterogeneity in the data might sufficiently be represented by fewer phenotypes (detailed results for the models with 2-5 phenotypes are reported in tables E2-E5 in the online supplementary material). The main characteristics of the five phenotypes are summarized below (details in tables 1 and 2). To simplify the discussion, each phenotype was given a summary label describing its most pertinent characteristics.

Phenotype A ("persistent cough"): Children with this phenotype typically suffered from cough apart from colds at both surveys. Wheeze ever was more common than in phenotype B but considerably less common than in phenotypes C, D and E. $FEV_{0.5}$ values tended to be slightly lower and bronchial responsiveness greater than in asymptomatic children.

Phenotype B ("transient cough"): Cough apart from colds occurred only in the first survey and wheeze ever was rarely reported. $FEV_{0.5}$ and bronchial responsiveness were comparable with asymptomatic children.

Phenotype C ("atopic persistent wheeze"): Attacks of wheeze were frequent in both surveys. Attacks occurred with and without colds and were commonly accompanied by shortness of breath. For almost a third of the children with this phenotype summer was the season with more frequent attacks in the second survey. Cough apart from colds and being woken at night by cough was common. Sensitization to at least one allergen was likely, $FEV_{0.5}$ values were typically lower and bronchial responsiveness greater than in asymptomatic children.

Phenotype D ("non-atopic persistent wheeze"): Attacks of wheeze were likely in both surveys though not as frequent as in phenotype C. Attacks tended to be accompanied by shortness of breath and occurred with and without colds. They were generally worse at night and, in the second survey, were more common in winter. Atopic sensitization was rare, $FEV_{0.5}$ similar and bronchial responsiveness greater than in asymptomatic children.

Phenotype E ("transient viral wheeze"): Attacks of wheeze tended to occur prior to the first survey or, if reported at the first survey, were infrequent. Attacks had subsided by the second survey. Wheeze tended to occur only with colds. $FEV_{0.5}$ was similar to that in asymptomatic children, bronchial responsiveness was slightly greater.

For each child in the sample membership probabilities were computed for each of the identified phenotypes. Children were then assigned to the phenotypes for which they had highest probability (phenotype clusters). For 271 children (85%) the highest membership probability was greater than 0.9 indicating clear membership, while for 9 children (3%) the highest membership probability was less than 0.6 indicating more ambiguous membership.

To investigate the relationship between phenotypes identified in the sequential steps of the analysis (models 1-5), we determined the number of children "flowing" from the phenotype clusters of a given model into the clusters of the subsequent model with one more phenotype (Figure 2). The phenotypes showed a high degree of stability across models. Children grouped to one phenotype at an early stage tended to be grouped together again at later stages. Thus four of the phenotypes of our five-phenotype model were essentially distinguished at earlier stages (phenotypes A and B by model 4 (clusters 4A and 4B) and phenotypes C and E by model 3 (3B and 3C)), with phenotype D appearing as the only "new" phenotype at the fifth stage.

Comparing prognosis across identified phenotypes

At age 8-13 years in 1998 (Figure 3, white columns) the prevalence of current wheeze was highest in phenotype cluster C ("atopic persistent wheeze") (37/52 = 71%), less in phenotype cluster D ("non-atopic persistent wheeze") (14/40 = 35%), followed by A ("persistent cough")

(21/84 = 25%) and E (“transient viral wheeze”) (8/34 = 24%) and lowest in B (“transient cough”) (7/72 = 10%) and in asymptomatics (17/158 = 11%). A similar pattern was found for the outcomes frequent wheeze (≥ 3 attacks) and use of bronchodilators.

We statistically tested for differences in the prevalence of the 4 prognostic endpoints between the phenotype clusters. We were interested in pair-wise comparisons between children with persistent cough (A) and asymptomatics and between the two cough phenotypes (A and B) because persistent coughers represent a novel group identified by this study (see discussion). It is still disputed whether children with chronic cough, or a subgroup of them, have a different probability to develop wheeze compared to asymptomatic children. We also tested for differences between the two more persistent wheeze phenotypes (C and D). In order to limit the problem of multiple testing, we did not perform more pair-wise comparisons. The Bonferroni-corrected significance level for these tests was 0.0042 (overall significance level divided by number of tests: 0.05/12). The outcomes at 8-13 years (Figure 3, white columns) tended to be more prevalent in cluster C than in D with significant differences for current wheeze ($p=0.001$) and for use of bronchodilators ($p=0.002$). Prognosis of asthma-related outcomes tended to be worse for phenotype cluster A (“persistent cough”) than for phenotype cluster B (“transient cough”) and asymptomatics, with significant differences for use of bronchodilators ($p<0.001$ and $p=0.001$ respectively). Prevalence of cough apart from colds at 8-13 years was higher in A (44%) than in B (18%; $p=0.001$) and in asymptomatics (12%; $p<0.001$).

In 2003, at 13-18 years (Figure 3, grey columns) prognostic differences between phenotype clusters remained qualitatively similar for all four outcomes. Marked differences, though not significant at the Bonferroni-corrected level, remained between C and D for current wheeze (56%; 31%; $p=0.038$), and inhaler use (65%; 36%; $p=0.013$). Prevalence of cough apart from colds again differed significantly between A (41%) and B (16%; $p=0.002$).

Discussion

This paper describes a novel approach to phenotype recognition in children with wheeze and cough using latent class analysis (LCA). By applying this method to data on respiratory symptoms and physiological measurements from a population-based childhood cohort, we identified three wheeze phenotypes and two cough phenotypes. These phenotypes were predictive of outcomes at school age and later childhood. What distinguishes these entities from previously used phenotypes is that they are derived directly from data, rather than defined a priori, and that they account for multiple disease dimensions.

LCA as multidimensional clustering technique

Historically, clinicians have refined diagnosis by resolving complex diseases into discrete clinically useful subsets. These subsets, which we refer to as disease phenotypes, provide a way of classifying patients into groups of individuals with similar disease characteristics. In this study phenotypes were treated as unknown and were derived from the observed heterogeneity in a sample of symptomatic children. The chosen technique, LCA, can be interpreted as a form of cluster analysis. It has, however, important advantages over algorithmic clustering techniques such as hierarchical or k-means clustering. First, it is based on a formal statistical model which can readily accommodate features measured in different modes (categorical, continuous or count variables). Second, the algorithm which is typically used to fit these models was designed to deal with missing values [27]. These models thus meet major challenges of real-life epidemiological and clinical data. Third, the resulting clusters are not rigid in the sense that each individual is assigned to just one class. Rather each individual can be assigned to various classes with differing probabilities. This soft form of classification more closely corresponds to the clinical situation where some patients have features common to more than one condition. In our sample, we found that the majority of children could clearly be classified into one of the phenotypes, that is with a high probability, but that for a minority of children there remained some ambiguity. A possible downside of our approach is that this method does not directly produce clear-cut diagnostic rules for the clinical setting. However, once phenotype definitions obtained by this technique are validated, for instance by application of the model to independent datasets, results can be translated into simplified diagnostic algorithms in a further stage.

LCA shares some limitations with other clustering techniques. First, the problem of determining the number of classes has not been completely resolved [11]. Different statistical criterion can be used to determine the number of classes, but may yield different results, as has been the case in our study. Second, some prior decisions need to be made, such as the type and number of variables to include. This method therefore also involves some degree of subjectivity, though considerably less than a priori phenotype definitions. In the present application another multivariate statistical method, multiple correspondence analysis, was used to assist variable selection and reduce the risk of subjective choices. The phenotypes identified are influenced by the range of data included. It is therefore necessary that all dimensions considered to be relevant for phenotype definition are represented by the included variables. As long as the same disease dimensions are included, results obtained by applying this approach to different cohorts should be comparable, even if the single variables representing these disease dimensions might differ (e.g. skin prick tests versus specific IgE measurements). In this analysis we have deliberately focused on clinical dimensions - signs, symptoms and physiological measurements - that is dimensions related to disease expression and not to disease causes. The reason for this was to keep the methodology as simple and transparent as possible at this early stage of research. Using appropriate adjustments to the statistical model, future applications may extend this approach to include important risk factors of wheezing disorders such as smoking.

The dataset used for this study which was obtained from an ongoing population-based cohort had a small sample size and a considerable proportion of missing values (12.8%). These problems are typical of clinical or epidemiological data. The dataset thus provided a suitable test bed for the new approach. Though only 11% of individuals had complete data for all variables, all 319 individuals contributed to the analysis. This highlights the advantage of using an estimation procedure which makes best use of all available information in spite of missing values. The fact that the response rates at survey 1 (1422/1650 = 86%) and survey 2 (488/795 = 61%) were not 100% might have induced some selection bias. This will mainly have affected the prevalence of identified phenotypes within our sample, but is less likely to have influenced the type of phenotypes found.

A further limitation of our study sample might have been the considerable age spread of the children at the time of data collection. The probability of observing certain features such as atopy or “wheeze ever” changes naturally with age. We partially accounted for this by including age in our model which allowed for a narrower age spread within phenotypes.

Phenotypes of wheeze

The model distinguished the phenotypes “transient viral wheeze” (phenotype E) related to colds and affecting mainly non-atopic children, and “atopic persistent wheeze” (phenotype C) associated with multiple triggers and atopy. This suggests that the previously proposed categorizations ‘transient’ and ‘persistent wheeze’ [2, 3, 28], and ‘viral’ and ‘multiple trigger wheeze’ [9, 29, 30] might reflect a single phenotypic dichotomy. Phenotypes E and C appear to reconcile the discrepancies between these two sets of labels. Children with the “atopic persistent wheeze” phenotype were mostly atopic, had the highest levels of bronchial responsiveness, lowest lung function and poorest prognosis, which agrees with findings from other groups [3, 31]. Children with the “transient viral wheeze” phenotype were generally non-atopic and had normal lung function. This matches findings from the German Multicentre

Asthma Study [32] but contrasts with reports from Tucson describing impaired lung function both in infancy and at early school age in early transient wheezers [3].

A third phenotype of wheeze (phenotype D in Fig 2) was labelled “non-atopic persistent wheeze” and was characterized by a low rate of atopy, similar to the phenotype labelled “non-atopic wheeze” by the Tucson group [28, 33]. A low rate of atopy and the winter season predominance distinguished this phenotype from the atopy-associated phenotype. It is known from experimental studies that a non-atopic form of viral wheeze may persist in mild form into adult life [34]. Phenotype D also shares features with what has been described as “intrinsic asthma” in adult respiratory medicine [35]. No evidence was found for a distinct “late onset” phenotype characterized by wheeze reported only in the second survey [2, 3, 28]. The application of LCA to the present dataset therefore provided support for a) the distinction between transient and persistent wheeze, recognizing that the former is associated with viral infections and the latter with other triggers, and for b) the existence of a third form of wheeze which is non-atopic but largely persistent.

Phenotypes of cough

One of the two cough phenotypes which our model identified (phenotype A) was associated with reduced lung function, increased bronchial responsiveness and a significantly higher risk of later wheeze compared to asymptomatics and to children belonging to the other cough phenotype (Figure 3). The statistical model therefore appears to have identified, within the large group of children with non-specific cough, a group which exhibits features of a condition called “cough-variant asthma”. It is clear that “lumping” together all children with chronic cough under this term leads to an over-diagnosis of asthma [8]. The present multidimensional approach might help to single out a subgroup of children who might indeed profit from asthma treatment.

Implications for research and clinical practice

Reliable phenotype definitions are important for research and clinical practice. They are useful for describing the natural history of the disease and for studying underlying mechanisms and the role of environmental and genetic factors. In the clinical setting, the ability to allocate children to phenotypes allows informed counselling of parents and is a prerequisite for phenotype-specific treatment [5-7]. More accurate phenotype definitions might also help to explain seemingly conflicting results in time trends and international prevalence of asthma [20, 36].

In all these settings, phenotypes are useful only if they reflect true disease entities. Statistical techniques which are designed to detect the structures underlying multivariate data, such as LCA, have the potential to identify such phenotypes. But, because these methods are exploratory, it is important to validate the resulting phenotypes. In the present study, recent outcome data were used to provide support for phenotypes identified from early symptoms and physiological measurements. Identifying similar phenotypes using independent data sets is an additional necessary step for validating these findings. Further development of this approach and application to other cohorts should help increase our understanding of phenotypic variability not only in childhood respiratory disorders but also in adult obstructive airway diseases, where phenotype definition is an equally important issue [10].

Acknowledgements
The authors thank all the children and their parents for participating in the study, Tony Davis, Business Manager, Children's Directorate, Leicester City West Primary Care Trust for his assistance with the Leicestershire Child Health Database, and Matthias Egger, Philip Latzin, Marie-Pierre Strippoli and Martin Brinkhof for their valuable comments on the manuscript. The authors also thank Jonathan Grigg who helped finance the follow-up survey in 2003 via a grant from the UK Department of Health (number 0020014).

References

1. Grigg J, Silverman M: Wheezing disorders in young children: one disease or several phenotypes? *In*: Frey U, Gerritsen J, eds. Respiratory diseases in infants and children. vol. 37: Eur Respir Mon, 2006; pp. 153-169.
2. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and wheezing in the first six years of life. *N Engl J Med* 1995; 332: 133-138.
3. Morgan WJ, Stern DA, Sherrill DL, *et al*. Outcome of Asthma and Wheezing in the First Six Years of Life: Follow-up through Adolescence. *Am J Respir Crit Care Med* 2005; 172: 1253-1258.
4. Bel EH. Clinical phenotypes of asthma. *Curr Opin Pulm Med* 2004; 10: 44-50.
5. Bush A. Phenotype specific treatment of asthma in childhood. *Paediatr Respir Rev* 2004; 5: S93-101.
6. Gold DR, Fuhlbrigge AL. Inhaled corticosteroids for young children with wheezing. *N Engl J Med* 2006; 354: 2058-2060.
7. Brand P, Bisgaard H, Baraldi E, *et al*. Definition, diagnosis and treatment of wheezing disorders in preschool children - an evidence based approach. European Respiratory Society Task Force Report. *Eur Respir J* submitted;
8. de Jongste JC, Shields MD. Cough 2: Chronic cough in children. *Thorax* 2003; 58: 998-1003.
9. Silverman M, Grigg J, Mc Kean M: Virus-induced wheeze in young children - A separate disease? *In*: Johnston S, Papadopoulos N, eds. Respiratory infections in allergy and asthma. New York: Marcel Dekker, 2002; pp. 427-471.
10. Wardlaw AJ, Silverman M, Siva R, Pavord ID, Green R. Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clin Exp Allergy* 2005; 35: 1254-1262.
11. McLachlan G, Peel D. Finite Mixture Models. New York: John Wiley & Sons; 2000.
12. Kohlmann T, Formann AK: Chapter 33: Using Latent Class Models to Analyze Response Patterns in Epidemiologic Mail Surveys. *In*: Rost J, Langeheine R, eds. Applications of Latent Trait and Latent Class Models in the Social Sciences. Münster, New York, München, Berlin: Waxmann, 1997; pp. 345-352.
13. Dunn KM, Jordan K, Croft PR. Characterizing the course of low back pain: a latent class analysis. *Am J Epidemiol* 2006; 163: 754-761.
14. Croudace TJ, Jarvelin MR, Wadsworth ME, Jones PB. Developmental typology of trajectories to nighttime bladder control: epidemiologic application of longitudinal latent class analysis. *Am J Epidemiol* 2003; 157: 834-842.
15. Kuehni CE, Brooke AM, Strippoli M-PF, Spycher BD, Davis A, Silverman M. Cohort profile: The Leicester Respiratory Cohorts. *Int J Epidemiol* 2007; 36: 977-985.
16. Luyt DK, Burton PR, Simpson H. Epidemiological study of wheeze, doctor diagnosed asthma, and cough in preschool children in Leicestershire. *BMJ* 1993; 306: 1386-1390.
17. Brooke AM, Lambert PC, Burton PR, Clarke C, Luyt DK, Simpson H. The natural history of respiratory symptoms in preschool children. *Am J Respir Crit Care Med* 1995; 152: 1872-1878.
18. Brooke AM, Lambert PC, Burton PR, Clarke C, Luyt DK, Simpson H. Recurrent cough: Natural history and significance in infancy and early childhood. *Pediatr Pulmonol* 1998; 26: 256-261.
19. Kuehni CE, Brooke AM, Silverman M. Prevalence of wheeze during childhood: Retrospective and prospective assessment. *Eur Respir J* 2000; 16: 81-85.
20. Kuehni CE, Davis A, Brooke AM, Silverman M. Are all wheezing disorders in very young (preschool) children increasing in prevalence? *Lancet* 2001; 357: 1821-1825.
21. Nystad W, Samuelsen SO, Nafstad P, Edvardsen E, Stensrud T, Jaakkola JJ. Feasibility of measuring lung function in preschool children. *Thorax* 2002; 57: 1021-1027.

22. Wilson NM, Phagoo SB, Silverman M. Use of transcutaneous oxygen tension, arterial oxygen saturation, and respiratory resistance to assess the response to inhaled methacholine in asthmatic children and normal adults. *Thorax* 1991; 46: 433-437.
23. Greenacre MJ. Theory and applications of correspondence analysis. London: Academic Press; 1984.
24. Hunt L, Jorgensen M. Mixture model clustering using the MULTIMIX program. *Aust N Z J Stat* 1999; 41: 153-171.
25. Chinn S. Methodology of bronchial responsiveness. *Thorax* 1998; 53: 984-988.
26. Hunt L, Jorgensen M. Mixture model clustering for mixed data with missing information. *Comput Stat Data Anal* 2003; 41: 429-440.
27. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Statist Soc Ser B* 1977; 39: 1-38.
28. Taussig LM, Wright AL, Holberg CJ, Halonen M, Morgan WJ, Martinez FD. Tucson Children's Respiratory Study: 1980 to present. *J Allergy Clin Immunol* 2003; 111: 661-675.
29. Silverman M, Wilson N. Asthma--time for a change of name? *Arch Dis Child* 1997; 77: 62-64.
30. Silverman M. Out of the mouths of babes and sucklings: lessons from early childhood asthma. *Thorax* 1993; 48: 1200-1204.
31. Ross S, Godden DJ, Abdalla M, *et al*. Outcome of wheeze in childhood: the influence of atopy. *Eur Respir J* 1995; 8: 2081-2087.
32. Lau S, Illi S, Sommerfeld C, *et al*. Transient early wheeze is not associated with impaired lung function in 7-yr-old children. *Eur Respir J* 2003; 21: 834-841.
33. Martinez FD. Development of wheezing disorders and asthma in preschool children. *Pediatrics* 2002; 109: 362-367.
34. McKean MC, Leech M, Lambert PC, Hewitt C, Myint S, Silverman M. A model of viral wheeze in nonasthmatic adults: symptoms and physiology. *Eur Respir J* 2001; 18: 23-32.
35. Kroegel C, Jager L, Walker C. Is there a place for intrinsic asthma as a distinct immunopathological entity? *Eur Respir J* 1997; 10: 513-515.
36. Beasley R, Ellwood P, Asher I. International patterns of the prevalence of pediatric asthma the ISAAC program. *Pediatr Clin North Am* 2003; 50: 539-553.

TABLE 1. Objective features in the five phenotypes of chronic cough and wheeze and an asymptomatic control group

Phenotype	A	B	C	D	E	Asymptomatic[#]
n[¶]	97	82	58	47	35	169
Sex						
Female	0.56	0.52	0.51	0.51	0.27	0.49
Male	0.44	0.48	0.49	0.49	0.73	0.51
Age in 1990						
0 to 2 yrs	0.39	0.54	0.30	0.61	0.36	0.51
3 to 5 yrs	0.61	0.46	0.70	0.39	0.64	0.49
Skin prick tests						
All negative	0.81	0.84	0.30	0.91	0.78	0.89
At least one positive	0.19	0.16	0.70	0.09	0.22	0.11
$FEV_{0.5}$ (z-scores)[+]						
Mean	-1.59	-1.18	-1.80	-1.47	-1.09	-1.33
SD	1.41	1.05	1.41	0.57	0.96	1.47
Bronchial responsiveness (PC20tc-PO_2 g/L)[§]						
Geometric mean	2.42	2.75	1.26	2.32	2.48	3.82
IQR	1.4-4.1	1.5-5.2	0.61-2.6	1.4-3.9	1.4-4.3	2.7-6.2

Data are probabilities for categorical variables, and means with standard deviations (SD) or interquartile ranges (IQR) for continuous variables as estimated by the finite mixture model. For example, children with phenotype E had a probability of 22% to have a positive skin prick test, while this is only 9% for phenotype D.

[#]: Children reporting no cough apart from colds and no wheeze ever in both early surveys (1990, 1992-4). These subjects were not included in the sample used for model estimation.

[¶]: Number of children assigned to phenotype.

[+]: Sex and weight adjusted z-scores [21]. The negative mean score in asymptomatic children probably reflects a systematic difference between our sample and the reference population which is irrelevant for this analysis.

[§]: Values based on antilog transformation of model estimated mean and IQR for log PC_{20} [25].

TABLE 2. Probabilities of symptoms at the two first surveys in the five phenotypes

Phenotype		**A**		**B**		**C**		**D**		**E**	
n[#]		97		82		58		47		35	
Survey		90	92-4	90	92-4	90	92-4	90	92-4	90	92-4
Wheeze ever	Yes	0.24	0.26	0.09	0.15	0.75	0.91	0.79	0.90	0.91	0.73
Frequency of attacks[¶+]	0	0.07	0.12	0.00	0.10	0.00	0.23	0.00	0.44	0.63	0.69
	1 to 2	0.11	0.05	0.06	0.02	0.16	0.31	0.51	0.31	0.28	0.05
	≥ 3	0.06	0.09	0.04	0.02	0.59	0.38	0.28	0.15	0.00	0.00
Attacks with shortness of breath[+]	No	0.11	0.14	0.06	0.02	0.02	0.61	0.23	0.45	0.48	0.64
	Yes	0.13	0.12	0.04	0.13	0.73	0.30	0.56	0.45	0.43	0.09
Triggers of wheeze[+]	Only colds	0.18	0.14	0.06	0.15	0.00	0.00	0.40	0.43	0.64	0.60
	Colds and other	0.06	0.12	0.04	0.00	0.75	0.91	0.39	0.46	0.28	0.13
Season with most frequent attacks[+]	Indifferent	0.18	0.08	0.09	0.15	0.43	0.29	0.41	0.08	0.83	0.73
	Winter	0.04	0.13	0.00	0.00	0.22	0.32	0.39	0.75	0.08	0.00
	Summer	0.02	0.04	0.00	0.00	0.09	0.31	0.00	0.06	0.00	0.00
Time of day with worse attacks[+]	Indifferent	0.08	0.07	0.00	0.04	0.23	0.18	0.23	0.00	0.64	0.46
	Night	0.16	0.19	0.08	0.09	0.47	0.61	0.46	0.80	0.28	0.27
	Day	0.00	0.00	0.01	0.02	0.04	0.12	0.10	0.09	0.00	0.00
Wakened by cough at night	Yes	0.32	0.69	0.79	0.19	0.82	0.63	0.31	0.46	0.06	0.05
Cough (RC = No cough)	Only with colds	0.29	0.00	0.00	0.85	0.40	0.38	0.72	0.83	0.65	0.67
	Also apart from colds	0.63	1.00	1.00	0.00	0.49	0.53	0.16	0.00	0.03	0.00

Legend:

0.00-0.25	0.25-0.50	0.50-0.75	0.75-1.00
□	▒	▓	█

RC: residual category (omitted category with which the probabilities sum to one)

Data are presented as probabilities. For instance children with phenotype C had a probability of 75% and 91% to have attacks of wheeze also apart from colds at the first and second survey respectively.

[#]: Number of children assigned to phenotype

[¶]: In the past 12 months

[+]: Questions were asked only to those who reported wheeze ever in the respective survey. The RC for these variables is “No” to wheeze ever.

Figure legends:

Figure 1: Flow diagram of study subjects. Survey 1 took place in 1990 and survey 2, on a nested sample, in 1992-4 in conjunction with laboratory studies.

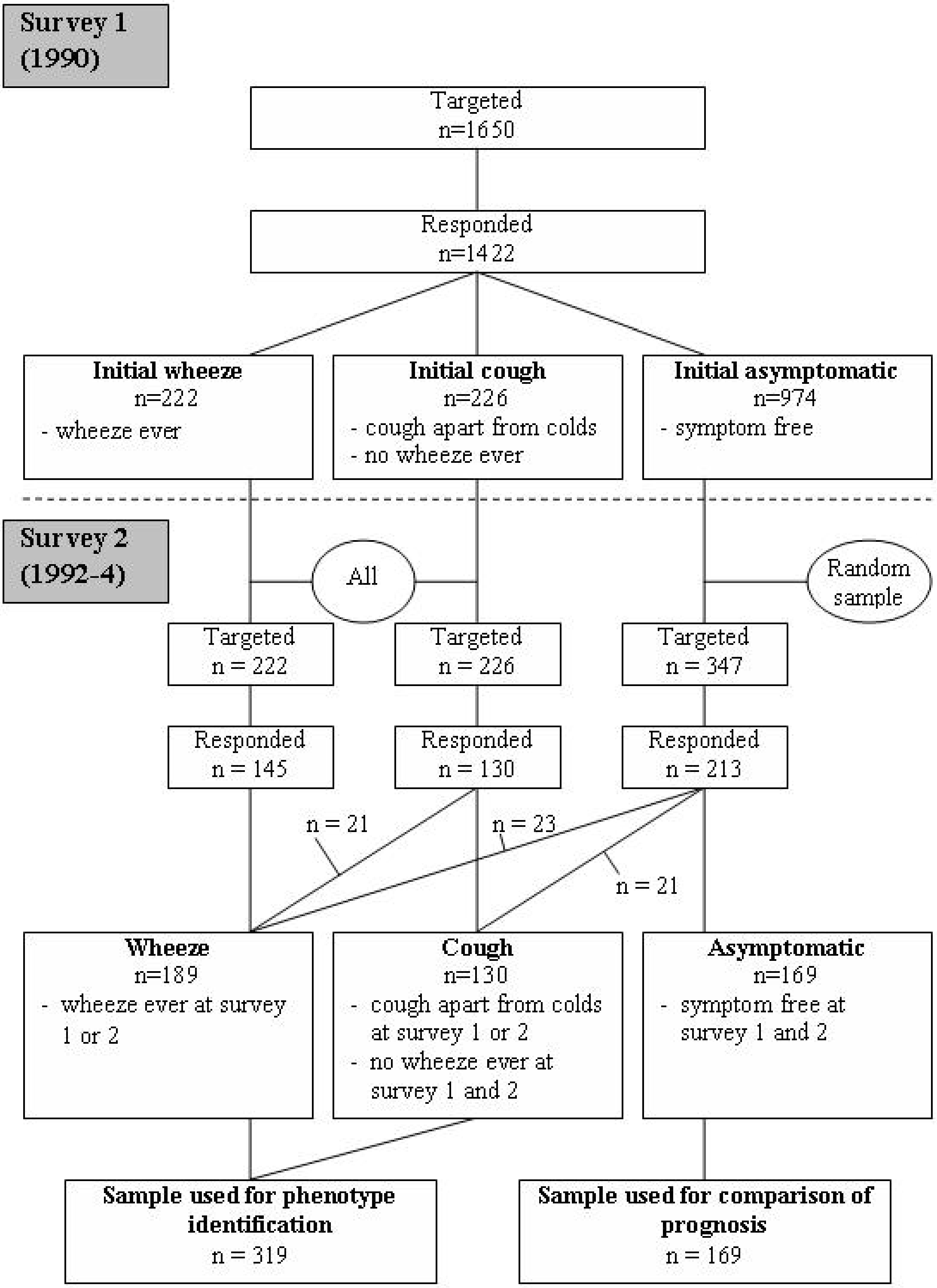

Survey 1 (1990)
Targeted n=1650
Responded n=1422
Initial wheeze n=222 - wheeze ever
Initial cough n=226 - cough apart from colds - no wheeze ever
Initial asymptomatic n=974 - symptom free
Survey 2 (1992-4)
All
Random sample
Targeted n = 222
Targeted n = 226
Targeted n = 347
Responded n = 145
Responded n = 130
Responded n = 213
n = 21
n = 23
n = 21
Wheeze n=189 - wheeze ever at survey 1 or 2
Cough n=130 - cough apart from colds at survey 1 or 2 - no wheeze ever at survey 1 and 2
Asymptomatic n=169 - symptom free at survey 1 and 2
Sample used for phenotype identification n = 319
Sample used for comparison of prognosis n = 169

Figure 2: Formation of phenotype clusters as the number of phenotypes in the mixture model was increased stepwise. Dark shaded boxes in a given layer represent clusters of children corresponding to the disease phenotypes identified by a given model. Models are labelled according to the number of phenotypes included in the model. Box widths are proportional to numbers of children in the respective clusters. Hatched parallelograms connecting adjacent

layers represent the numbers children (proportional to horizontal width) common to any two connected clusters. For each model clusters are labelled #A, #B, etc. from left to right with '#' representing the number of phenotypes in the model. In the text the phenotypes of model 5 are referred to using the letters shown below the figure.

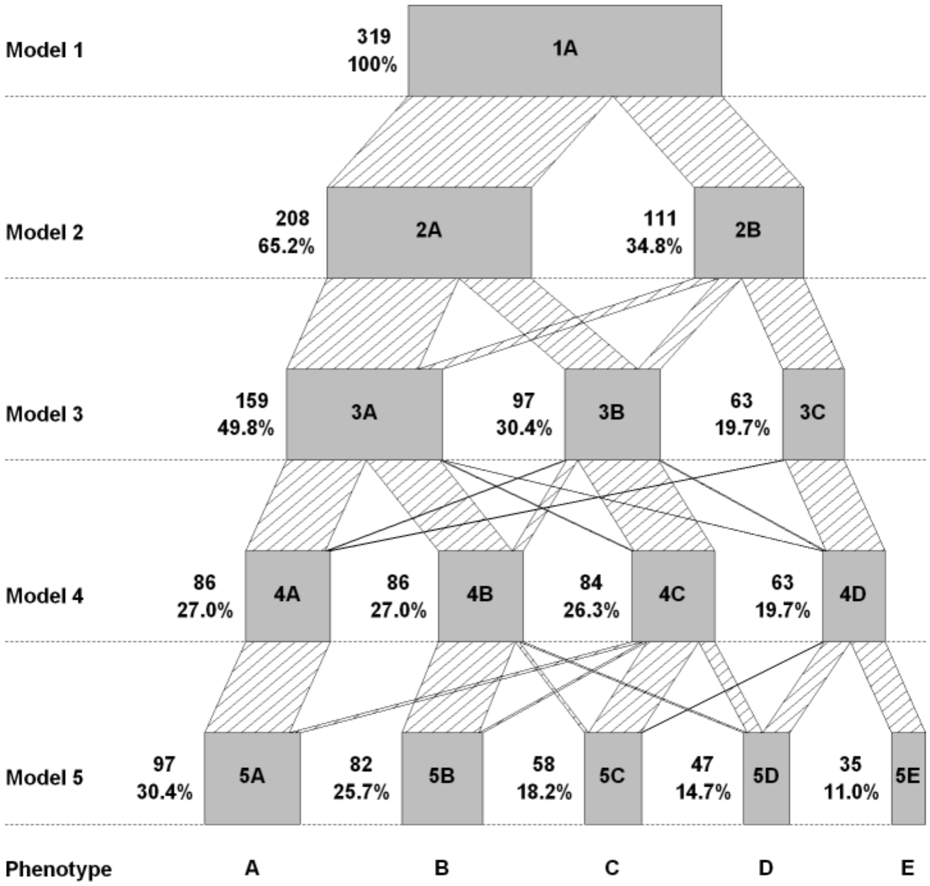
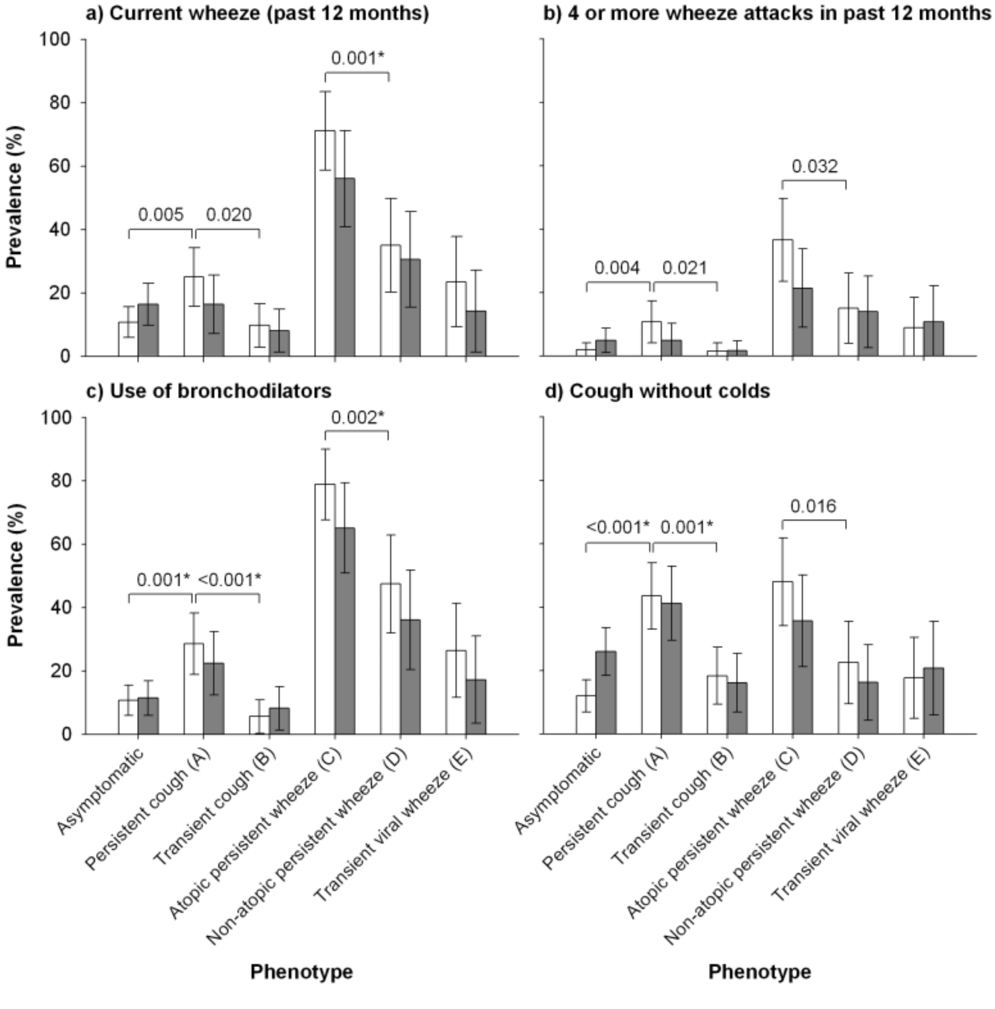


Figure 3: Prognostic outcomes 5 and 10 years later for the five phenotype clusters and an asymptomatic control group. Columns represent the prevalence of outcomes at age 8-13 years (☐) and 13-18 years (■). Error bars indicate 95% confidence intervals. P-values (two sided Fisher's exact test) are shown for certain pair-wise comparisons of outcomes at 8-13 years. Significant values at the Bonferroni-corrected α level of 0.0042 are marked with an asterisk. Phenotypes were subjectively assigned labels based on their most important characteristics.


a) Current wheeze (past 12 months)
b) 4 or more wheeze attacks in past 12 months
c) Use of bronchodilators
d) Cough without colds
Prevalence (%)
0
20
40
60
80
100
0.005
0.020
0.001*
0.004
0.021
0.032
0.001*
<0.001*
0.002*
<0.001*
0.001*
0.016
Asymptomatic
Persistent cough (A)
Transient cough (B)
Atopic persistent wheeze (C)
Non-atopic persistent wheeze (D)
Transient viral wheeze (E)
Phenotype