Using latent class growth analysis to identify childhood wheeze phenotypes in an urban birth cohort

Qixuan Chen, PhD *; Allan C. Just, PhD †,‡; Rachel L. Miller, MD †,‡,§,¶; Matthew S. Perzanowski, PhD †,‡; Inge F. Goldstein, DrPH †,‖; Frederica P. Perera, DrPH †,‡; and Robin M. Whyatt, DrPH †,‡

* Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York
† Columbia Center for Children's Environmental Health, Mailman School of Public Health, Columbia University, New York, New York
‡ Department of Environmental Health Science, Mailman School of Public Health, Columbia University, New York, New York
§ Division of Pulmonary, Allergy, Critical Care, Department of Medicine, Columbia University College of Physicians and Surgeons, New York, New York
¶ Department of Pediatrics, Columbia University College of Physicians and Surgeons, New York, New York
‖ Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York

ARTICLE INFO

ABSTRACT

Background: To advance asthma cohort research, we need a method that can use longitudinal data, including when collected at irregular intervals, to model multiple phenotypes of wheeze and identify both time-invariant (eg, sex) and time-varying (eg, environmental exposure) risk factors.

Objective: To demonstrate the use of latent class growth analysis (LCGA) in defining phenotypes of wheeze and examining the effects of causative factors, using repeated questionnaires in an urban birth cohort study.

Methods: We gathered repeat questionnaire data on wheeze from 689 children ages 3 through 108 months ($n = 7{,}048$ questionnaires) and used LCGA to identify wheeze phenotypes and model the effects of time-invariant (maternal asthma, ethnicity, prenatal environmental tobacco smoke, and child sex) and time-varying (cold/influenza [flu] season) risk factors on prevalence of wheeze in each phenotype.

Results: LCGA identified four wheezing phenotypes: never/infrequent (47.1%), early-transient (37.5%), early-persistent (7.6%), and late-onset (7.8%). Compared with children in the never/infrequent phenotype, maternal asthma was a risk factor for the other 3 phenotypes; Dominican versus African American ethnicity was a risk factor for the early-transient phenotype; and male sex was a risk factor for the early-persistent phenotype. The prevalence of wheeze was higher during the cold/flu season than otherwise among children in the early-persistent phenotype ($P = .08$).

Conclusion: This is the first application of LCGA to identify wheeze phenotypes in asthma research. Unlike other methods, this modeling technique can accommodate questionnaire data collected at irregularly spaced age intervals and can simultaneously identify multiple trajectories of health outcomes and associations with time-invariant and time-varying causative factors.

© 2012 American College of Allergy, Asthma & Immunology. Published by Elsevier Inc. All rights reserved.

Introduction

Although wheeze is a common respiratory symptom,[1,2] distinct patterns exist during childhood.[3,4] Martinez and colleagues[3] defined 4 wheeze phenotypes (no wheezing, transient-early, late-onset, and persistent wheezing) based on whether the child had reported lower respiratory tract illness with wheezing during the first 3 years of life and whether they were wheezing at age 6 years.[3] This classification has several limitations: (1) children are classified into a single phenotype without estimation of uncertainty; (2) it does not allow simultaneous modeling of phenotypic patterns with causative factors that predict phenotypes or prevalence of wheeze at each age; (3) it cannot be applied to other birth cohort studies with wheeze reports at different ages; and (4) it does not allow the comparison of phenotypes across cohorts.

Recently, longitudinal latent class analysis (LLCA) has been used to characterize distinctive wheeze phenotypes in several asthma cohorts.[5–8] LLCA is a statistical method that can be used to cluster individuals into a number of latent classes (phenotypes) based on the pattern of response to wheeze questions at discrete time points. It avoids the need to define phenotypes by the onset of wheeze at prespecified ages. However, it still has 2 limitations: it requires the

wheeze questions to be collected at regularly spaced intervals; and it does not allow modeling the effect of time-varying causative factors (eg, environmental exposure at each age) on prevalence of wheeze.

The limitations of LLCA can be overcome by the use of latent class growth analysis (LCGA),[9,10] another type of latent class analysis that has not been applied to asthma research previously. In LCGA, age at questionnaire collection is treated as a continuous variable, and the trajectory of the development of wheeze is estimated as a continuous function of age. Treating age at questionnaire collection as a continuous variable is especially useful, because few longitudinal studies collect questionnaires at exactly the same child ages. Furthermore, LCGA allows not only the inclusion of time-invariant causative factors to predict phenotypes of wheeze, but also the inclusion of time-varying causative factors measured at each age to estimate the association of these risk factors and prevalence of wheeze in each phenotype.

In this article, we used data from the Columbia Center for Children's Environmental Health (CCCEH) birth cohort study to illustrate the application of LCGA in defining wheezing phenotypes. We explored common time-invariant causative factors (maternal asthma and ethnicity, prenatal exposure to environmental tobacco smoke, and child's sex) that may predict phenotypes of wheeze and a classical time-varying causative factor (cold/flu season) that may predict prevalence of wheeze at each age in each phenotype. Results were also compared with those using LLCA.

Methods

Study participants

Participants were from the CCCEH longitudinal birth cohort study that enrolled pregnant nonsmoking Dominican and African American women free of diabetes, hypertension, and known human immunodeficiency virus, and recruited from two prenatal clinics in Northern Manhattan (n = 727) as detailed previously.[11,12] Thirty-eight enrolled mothers dropped out after delivery and did not answer any wheeze questionnaires; thus, n = 689 participants were included in the current study. A signed informed consent was obtained in accordance with the Columbia University Institutional Review Board.

Definition of variables

Maternal asthma (yes/no) and ethnicity (Dominican/African American) were classified based on maternal report. Prenatal exposure to environmental tobacco smoke (yes/no) was classified as reporting a smoker in the home on the prenatal questionnaire or a maternal or cord blood cotinine measure over 15 ng/mL where available (n = 663/689). Repeated questionnaires were administered in person to the mother at up to 7 ages (6, 12, 24, 36, 60, 84, and 108 months) and by telephone (94–99% with the mother) at up to 8 ages (3, 9, 15, 18, 21, 30, 48, and 72 months). Questionnaires were administered in English and Spanish and asked, "In the past 3 months has your child had wheezing or whistling in the chest?" Questionnaires administered between September 1 and March 31 were defined as occurring during the cold/flu season. Questionnaires (n = 7,048) in this analysis were collected between October 1998 and May 2011.

Statistical analysis

The statistical analysis was conducted in two steps. First, the number of phenotypes (latent classes) was identified using LCGA and compared with LLCA. In the LCGA model, the number of phenotypes was determined by selecting the model with a minimum value of Bayesian Information Criterion,[10,13] an information criterion that combines goodness of fit and parsimony of the model. Cubic trajectories were used in all classes of the LCGA model.

Table 1
The number of completed questionnaires and the proportion of reported wheeze at each age in the full cohort (n = 689)

Age (months)	Method of administration	Number	Report of wheeze (%)
3	By phone	570	12.6
6	In person	609	23.5
9	By phone	480	17.9
12	In person	607	20.4
15	By phone	386	14.5
18	By phone	420	12.6
21	By phone	398	11.8
24	In person	561	15.5
30	By phone	395	12.7
36	In person	558	11.1
48	By phone	417	9.8
60	In person	545	12.3
72	By phone	396	8.3
84	In person	406	11.8
108	In person	300	11.0

Because continuous age is allowed in LCGA, the decimal months of age at each questionnaire were calculated and used. In the LLCA model, the number of phenotypes were identified using bootstrap likelihood ratio tests,[10] which were used to compare sequentially models with T versus T + 1 classes. In each phenotype, the wheeze prevalence was estimated independently at each of the 15 questionnaire ages in the LLCA model. As a second step, the chosen models were reestimated by introducing covariates, including maternal asthma and ethnicity, prenatal exposure to environmental tobacco smoke, and child's sex, to predict phenotypes. The probability of belonging to a phenotype depends on the values of the covariates through a multinomial logistic regression with the phenotype membership as the outcome variable. In the LCGA model, the cold/flu season status at each questionnaire also was included in the model to examine the effect of a time-varying risk factor on the prevalence of wheeze at each age. SAS 9.2 (SAS Institute Inc., Cary, North Carolina) was used for both the LCGA and the LLCA models, where LCGA was performed using the TRAJ procedure,[14,15] and LLCA was performed using the LCA procedure.[16–18]

Results

Descriptive statistics

Of the 689 children, 22.5% of mothers had asthma, 64.4% of mothers were Dominican, 35.6% of mothers were African American, 48.2% were boys, and 35.0% were exposed to prenatal environmental tobacco smoke. The reported wheezing prevalence at each collection ranged from 8.3% to 23.5%, and the prevalence of wheeze was higher at young ages (Table 1).

Wheezing phenotypes

The Bayesian Information Criterion selected a 4-class LCGA model (Fig 1). For each phenotype, the probability of wheezing was estimated as a cubic function of age adjusting for cold/flu season, and the 95% confidence bands were plotted to represent the uncertainty in the estimation. In addition to the trajectory with low probability of wheeze at all ages (named "never/infrequent" phenotype with a prevalence of 47.1%), 2 trajectories exhibited a high probability of wheeze in the early years of life. We labeled the one that had a low probability of wheeze at the older ages "early-transient" (with a prevalence of 37.5%) and the other one "early-persistent" (with a prevalence of 7.6%). A final trajectory described the onset of wheeze at the older ages and was named "late-onset" wheezing phenotype (with a prevalence of 7.8%). The early-transient wheezing children experienced relatively increased prevalence of wheeze before 36 months, but their prevalence declined to be only slightly higher than that of the children in the never/infrequent class after-
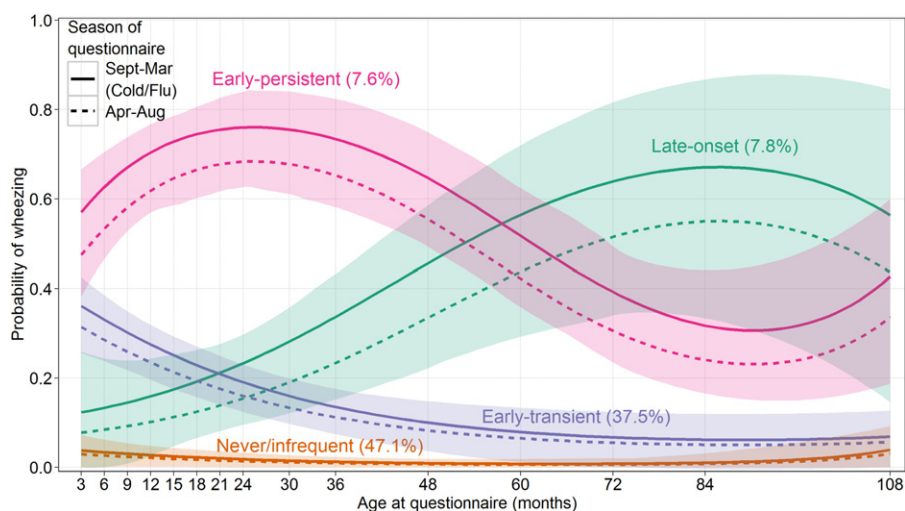
Fig. 1. Phenotypes of wheeze identified by the latent class growth analysis (LCGA) model using repeated questionnaire data from the CCCEH longitudinal birth cohort ($n = 689$).

wards. The early-persistent wheezing children had a high prevalence of wheeze before age 60 months, which decreased slightly by 108 months. For the children in the late-onset wheezing class, the prevalence of wheeze increased with age. Fewer questionnaires at ages older than 60 months led to wider confidence bands, indicating larger uncertainty in the estimation of wheeze prevalence at these ages. The prevalence of wheeze was higher during the 7 months of cold/flu season (denoted using solid curves) than otherwise (denoted using dashed curves), but this increase in prevalence was only marginally significant in the early-persistent wheezing group ($P = .08$) but not in other wheezing groups ($P > .1$).

Similar to the LCGA model, bootstrap likelihood ratio tests selected a 4-class LLCA model (Fig 2). The estimated proportion of children in each of the 4 phenotypes also is similar to that in the LCGA model. However, unlike the LCGA model, which estimated the prevalence of wheeze as a continuous function of age, LLCA estimated the prevalence of wheeze at the 15 discrete questionnaire ages in each class. The estimated prevalence of wheeze at the ages when the questionnaires were collected in person (denoted using *filled circles*) was in general higher than at the ages when the questionnaires were collected by phone (denoted using *hollow*

circles). This difference resulted in some fluctuation in each phenotypic pattern of wheeze. Furthermore, unlike the LCGA model, which estimated the prevalence of wheeze stratifying by cold/flu season, the LLCA model does not allow the modeling of time-varying risk factors. Thus, the prevalence of wheeze at each age could not be estimated separately for whether wheeze was reported in the cold/flu season.

Risk factors predicting phenotypes

Both the LCGA and LLCA models also examined time-invariant causative factors that predicted wheezing phenotypes simultaneously with the modeling of phenotypes. Figure 3 displays the odds ratio (OR) estimates with 95% confidence intervals (CIs) for each of the time-invariant causative factors by phenotype. LCGA showed that, compared with the children in the never/infrequent wheezing phenotype, maternal asthma (OR 2.6 [95% CI: 1.3–5.3]) and Dominican ethnicity versus African American (OR 2.4 [95% CI: 1.4–4.2]) were risk factors for the early-transient phenotype; maternal asthma (OR 5.7 [95% CI: 2.6–12.3]) and male sex (OR 4.1 [95% CI: 1.8–9.0]) were risk factors for the early-persistent phenotype; and maternal asthma (OR 3.8 [95% CI: 1.4–10.6]) was a risk factor for the late-onset phenotype. Prenatal exposure to environmental to-
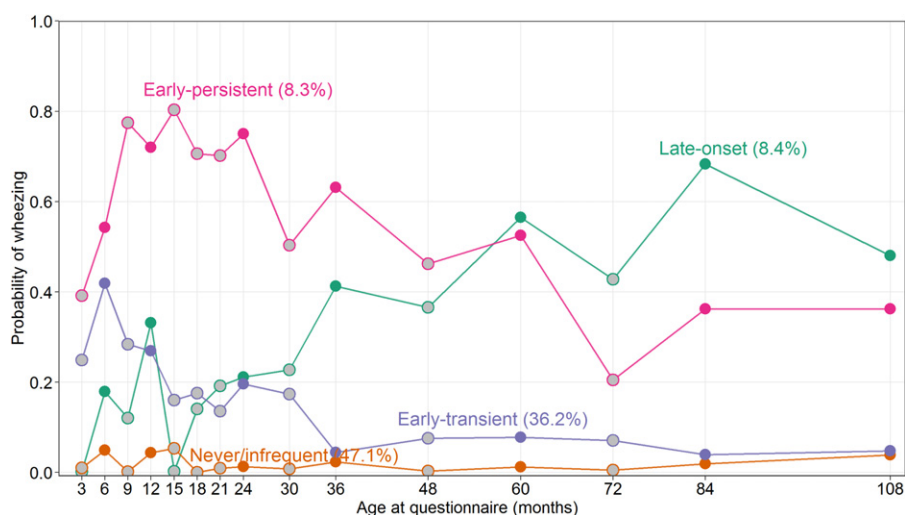

Fig. 2. Phenotypes of wheeze identified by the longitudinal latent class analysis (LLCA) model using repeated questionnaire data from the CCCEH longitudinal birth cohort ($n = 689$).
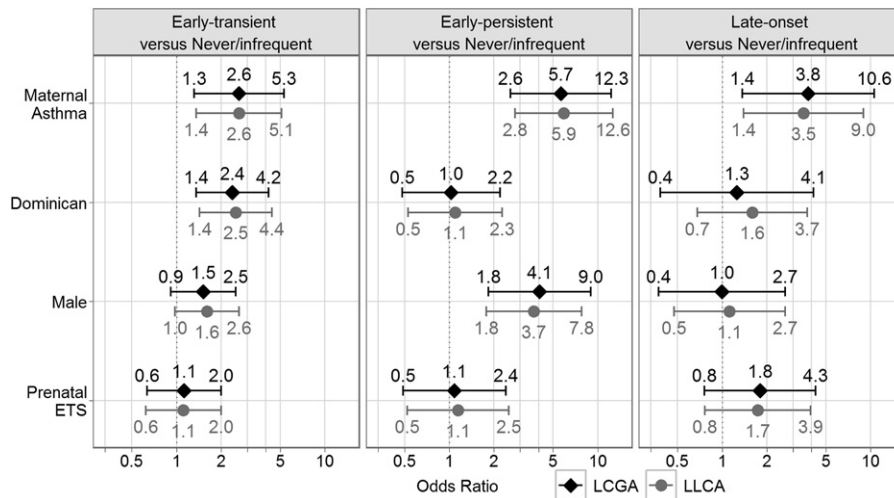
Fig. 3. OR estimates with 95% CI for the time-invariant causative factors—maternal asthma, ethnicity (Dominican/African American), child's sex, and prenatal environmental tobacco smoke (ETS)—by phenotype, with the never/infrequent phenotype as the reference.

bacco smoke was not a significant predictor for any phenotype. Similar odds ratios for these risk factors were identified using LLCA (Fig 3).

Sensitivity analysis

As reports of measurement of wheeze were missing in some questionnaires, especially when they were administered by telephone, LCGA was repeated among the children ($n = 431$) who did not have missing reports of wheeze for the first 5 questionnaires administered in person (ages 6, 12, 24, 36, and 60 months). The Bayesian Information Criterion identified the same 4 latent classes in this subsample with similar phenotypic patterns as in the full sample. The effects of the four time-invariant causative factors on predicting wheezing phenotypes are similar to the full sample, except that maternal asthma was no longer a significant risk factor for the late-onset phenotype (OR 3.1 [95% CI: 0.7–13.2]). See eFigure 1 and eFigure 2 for the profiles of the wheezing phenotypes and the effects of risk factors estimated from the subsample.

Discussion

A few studies have characterized phenotypes of wheeze during early childhood,[5,6] and many others have studied risk factors for developing predefined wheeze phenotypes or risk factors for predicting prevalence of wheeze at a single time point.[3,19] In this study, we demonstrated the use of LCGA in identifying phenotypes of wheeze and examining the effects of causative factors simultaneously in a data-driven manner, using repeated wheeze questionnaires during the first nine years of childhood in the CCCEH birth cohort. The application of the LCGA method is novel, because it allows us to simultaneously look for patterns of wheeze over time, time-invariant factors predicting phenotypic patterns, and time-varying factors predicting prevalence of wheeze at each age, using large amounts of questionnaire data in the same model. Although the odds ratio for time-invariant risk factors examined here were similar in the LCGA and LLCA models, the LCGA has the advantage of being able to include time-varying predictors in the model as well. We have demonstrated this by including information about whether the questionnaire at each age point was administered during the cold/flu season. As shown in Figure 1, a higher probability of wheeze during the cold/flu season for both early-persistent and late-onset wheeze phenotypes, although marginal significance was only observed in the early-persistent wheezing phenotype. Much less variability in the probability of wheeze during the cold/flu season was seen for the other 2 phenotypes. The ability to

include time-varying risk factors in the LCGA offers substantial advantages over other modeling techniques and may be particularly advantageous for research on environmental risk factors, such as ambient air pollution or pet exposure, which are likely to vary over time. For example, a previous study in the CCCEH cohort examined the relationship between cat ownership and the incidence of wheeze at ages 1, 2, 3, and 5, using 4 separate logistic regression models, and found differences in susceptibility to wheeze symptom depending on when a child was exposed.[20] The LCGA can be applied to remodel this relationship by treating the cat ownership as a time-varying covariate and using repeated measures of cat ownership and wheeze at multiple age points simultaneously. With the LCGA, we can examine not only the association between cat ownership and the development of wheeze over time, but also whether this association differs between different wheeze phenotypes. The association between the cold/flu season and the prevalence of wheeze was modeled here by simultaneously accounting for multiple unevenly spaced data collections, the heterogeneity of trajectories in patterns of wheezing over time, and the 4 time-invariant etiologic factors. LCGA also had advantages compared with LLCA in that age is included in the model as a continuous variable, whereas the LLCA method only allows discrete time points. Like most epidemiological longitudinal studies, each child's questionnaires were not collected at exactly the same ages for the targeted 15 data collection points in the CCCEH study, so that rounding of age was required in the use of the LLCA, and thus measurement errors were present.

Four distinct phenotypic patterns of wheeze were revealed by LCGA and, in keeping with previously described phenotypes, we named them never/infrequent, early-transient, early-persistent, and late-onset phenotype. An estimated one half of the children in the CCCEH cohort belonged to the never/infrequent phenotype; one third of children belonged to the early-transient phenotype; and only one sixth of children experienced more frequent wheezing symptoms throughout childhood, labeled as early-persistent or late-onset phenotypes. Our investigation of wheeze phenotypes in the CCCEH cohort validated the previous finding by Martinez and colleagues.[3] However, compared with Martinez's subjective phenotype definition, LCGA is novel in defining wheezing phenotypes in a data-driven manner that avoids the need to define the wheeze phenotypes by the onset of wheeze at some predefined age. This technique for defining phenotypes is advantageous because it allows the comparison of phenotypes across different populations with repeated measures of wheeze at different ages and can be

used to characterize individuals for prediction of future health outcomes based on their pattern across variable numbers of repeated assessments. The probability of phenotype membership for each child can be estimated and used in future studies to investigate the relationship between phenotype membership and other health outcomes, such as the development of asthma at later ages, use of asthma medication, and emergency room visits.

With both LLCA and LCGA methods, we were able to show differential causative factors for the identified phenotypes in the inner-city population–based CCCEH birth cohort. Children whose mothers had asthma were more likely to belong to any of the other 3 phenotypes than the never/infrequent phenotype, and the magnitude of the effect size was largest for the early-persistent phenotype and smallest for the early-transient phenotype. Although the findings of the association between maternal asthma and both early-persistent and late-onset phenotypes is consistent with reported research,[21–23] we did not expect maternal asthma to be associated with the early-transient phenotype, although others have reported on genetic variation associated with early wheezing caused by respiratory syncytial virus[24] and childhood environmental tobacco smoke exposure.[25] Boys had a higher chance of developing early-persistent wheezing phenotypes than girls, consistent with previous reports.[26] In addition, the early-transient wheezing phenotype was more commonly observed among children whose mothers were Dominican than African American.

We acknowledge that a limitation of this study is the relatively small number of questionnaires collected at ages after 60 months, because some of the children were not yet old enough to provide these data. Consequently, larger uncertainty in the prevalence of wheezing at ages after 72 months was observed, especially among the late-onset phenotype, which needs to be verified with more data in the future. As more children in the CCCEH study reach these ages, we will be able to estimate the probabilities of wheezing at these older ages with increased precision.

In conclusion, LCGA provides a flexible and useful tool in cohort research to study the associations of time-invariant and time-varying causative factors and health outcomes with heterogeneous trajectories measured at multiple time points. LCGA may be particularly advantageous when studying the effects of repeated environmental exposures and outcomes that vary over time through childhood.
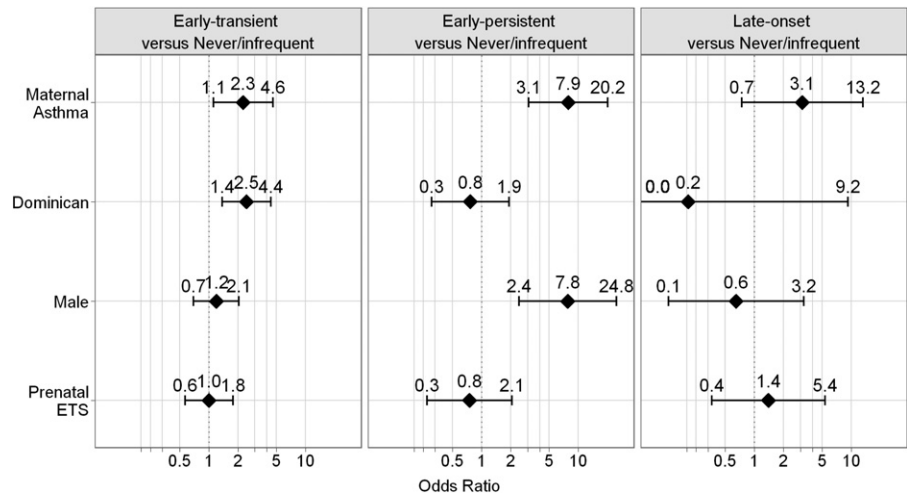
References

[1] Bloomberg, GR. Recurrent wheezing illness in preschool-aged children: Assessment and management in primary care practice. *Postgrad Med*. 2009;121: 48–55.

[2] Kurukulaaratchy RJ, Fenn M, Twiselton R, Matthews S, Arshad SH. The prevalence of asthma and wheezing illnesses amongst 10-year-old schoolchildren. *Respir Med*. 2002;96:163–169.

[3] 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.

[4] Kurukulaaratchy RJ, Fenn MH, Waterhouse LM, Matthews SM, Holgate ST, Arshad SH. Characterization of wheezing phenotypes in the first 10 years of life. *Clin Exp Allergy*. 2003;33:573–578.

[5] Savenije OE, Granell R, Caudri D, et al. Comparison of childhood wheezing phenotypes in 2 birth cohorts: alspac and piama. *J Allergy Clin Immunol*. 2011; 127:1505–1512 e1514.

[6] Henderson J, Granell R, Heron J, et al. Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood. *Thorax*. 2008;63:974–980.

[7] Spycher BD, Silverman M, Kuehni CE. Phenotypes of childhood asthma: Are they real? *Clin Exp Allergy*. 2010;40:1130–1141.

[8] Beath KJ, Heller GZ. Latent trajectory modelling of multivariate binary data. *Stat Model*. 2009;9:199–213.

[9] Nagin DS. Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychol Methods*. 1999;4:139–177.

[10] Jung T, Wickrama KAS. An introduction to latent class growth analysis and growth mixture modeling. *Social & Personality Psychology Compass*. 2008;2: 302–17.

[11] Perera FP, Rauh V, Tsai W-Y, et al. Effects of transplacental exposure to environmental pollutants on birth outcomes in a multiethnic population. *Environ Health Perspect*. 2003;111:201–205.

[12] Miller RL, Chew GL, Bell CA, et al. Prenatal exposure, maternal sensitization, and sensitization in utero to indoor allergens in an inner-city cohort. *Am J Respir Crit Care Med*. 2001;164:995–1001.

[13] Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6:461–464.

[14] Jones BL, Nagin DS, Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. *Sociol Method Res*. 2001;29:374–393.

[15] Jones BL, Nagin DS. Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociol Method Res*. 2007;35:542–571.

[16] Lanza ST, Dziak JJ, Huang L, Xu S, Collins LM. PROC LCA & PROC LTA user's guide (version 1.2.7). University Park: The Center, Penn State. 2011; Available at http://methodology.psu.edu. Accessed February 26, 2012.

[17] Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS procedure for latent class analysis. *Struct Equ Modeling*. 2007;14:671–694.

[18] Dziak JJ, Lanza ST, Xu S. LcaBootstrap SAS macro users' guide (version 1.1.0). University Park: The Methodology Center, Penn State. 2011; Available at http://methodology.psu.edu. Accessed February 26, 2012.

[19] Morgan WJ, Martinez FD. Risk factors for developing wheezing and asthma in childhood. *Pediatr Clin North Am*. 1992;39:1185–1203.

[20] Perzanowski MS, Chew GL, Divjan A, et al. Cat ownership is a risk factor for the development of anti-cat IgE but not current wheeze at age 5 years in an inner-city cohort. *J Allergy Clin Immunol*. 2008;121:1047–1052.

[21] Rusconi F, Galassi C, Corbo GM, et al. Risk factors for early, persistent, and late-onset wheezing in young children. Sidria collaborative group. *Am J Respir Crit Care Med*. 1999;160:1617–1622.

[22] Guill MF. Asthma update: Epidemiology and pathophysiology. *Pediatr Rev*. 2004;25:299–305.

[23] Smyth RL. Asthma: a major pediatric health issue. *Respir Res*. 2002;3 Suppl 1: S3–7.

[24] Hoffjan S, Ostrovnaja I, Nicolae D, et al. Genetic variation in immunoregulatory pathways and atopic phenotypes in infancy. *J Allergy Clin Immunol*. 2004;113: 511–518.

[25] Gilliland FD, Li YF, Dubeau L, et al. Effects of glutathione s-transferase m1, maternal smoking during pregnancy, and environmental tobacco smoke on asthma and wheezing in children. *Am J Respir Crit Care Med*. 2002;166:457–463.

[26] Almqvist C, Worm M, Leynaert B. Impact of gender on asthma in childhood and adolescence: A GA2LEN review. *Allergy*. 2008;63:47–57.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.anai.2012.02.016
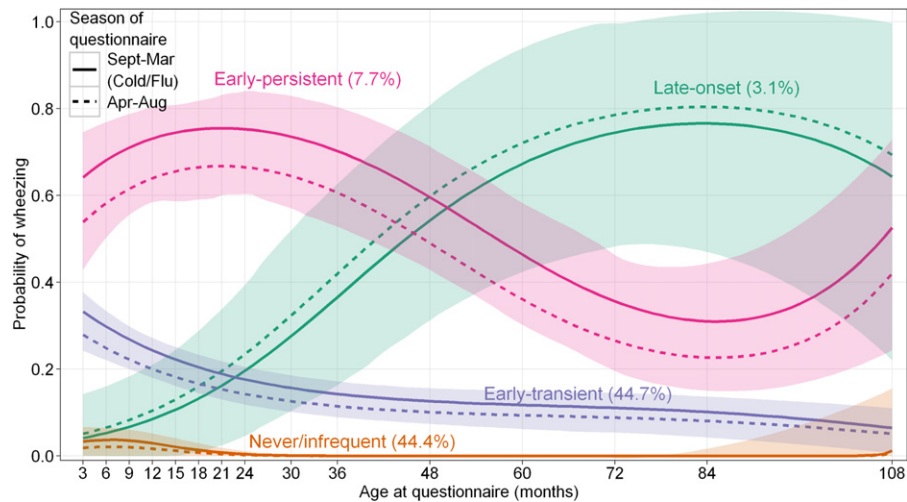
eFig. 1. Phenotypes of wheeze identified by the LCGA model using the subsample repeated questionnaire data among the 431 children who did not have missing reports of wheeze for the first 5 questionnaires administered in person from the CCCEH longitudinal birth cohort. The prevalence of wheeze in the past 3 months for each phenotype was estimated as a cubic function of age. Solid and dashed curves were used to denote the estimated functions of prevalence of wheeze during the 7 months of cold/flu season vs otherwise, respectively, stratified by phenotypes.

eFig. 2. Odds ratios estimates with 95% confidence intervals in the LCGA model for the 4 time-invariant risk factors by phenotype, with the never/infrequent wheezing phenotype as the reference group, using the subsample questionnaire data of the 431 children. The risk factors considered here include maternal asthma and ethnicity (Dominican vs African American), child's sex, and prenatal exposure to environmental tobacco smoke.