Imperial College London

Department of Population Health & Occupational Disease

THE APPLICATION OF BAYESIAN AND FREQUENTIST REGULARIZATION AND VARIABLE SELECTION METHODS FOR THE PREDICTION OF ASTHMA IN LATER CHILDHOOD

By

Silvia Colicino

Supervised by Prof Paul Cullinan, Dr Cosetta Minelli and Dr Alexandra Lewin

Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy of Imperial College London

Declaration

The work presented is my own work unless referenced or stated.

Silvia Colicino

November 2019

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgments

There are many people I would like to thank for their precious support and guidance over the last few years. Firstly, I would like to thank my supervisors Prof Paul Cullinan, Dr Cosetta Minelli and Dr Alexandra Lewin for the time and effort that they dedicated to offer constructive comments on the draft of chapters and their supervision and encouragement. Their professional supervision provided me with understanding of asthma prediction and gave me the strength to finish my PhD project.

Secondly, I would like to thank the Colt foundation and the STELAR consortium for their funding through the years.

Thirdly, I would also like to thank the STELAR children who participated in this research; none of this work would have been possible without their altruism, time and commitment.

Special thanks to my friends Matteo and Davide, I would have been lost without their motivation, humour, patience and love.

Lastly, I would like to thank my family who has always supported me in all decisions I have made. Even during difficult moments and when I doubted this project would ever come to an end, they provided me the encouragement to carry on.

Abstract

Asthma is a global health problem and among the most common chronic conditions in childhood. Several models were proposed to predict asthma in children, but their reproducibility in external populations was limited and none was developed to predict asthma in adolescence.

I conducted a systematic review of asthma predictive models validated in external populations; validation studies showed poorer predictive performances than development studies.

I developed predictive models for asthma between 15 and 20 years, using data from the Study Team for Early Life Asthma Research (STELAR) consortium of five UK asthma cohorts. For one of these cohorts, the Ashford study, I developed an questionnaire to collect follow-up information when study subjects were age 20 years.

I harmonised 41 variables across the STELAR cohorts, 39 of which were used as candidate predictors to develop predictive models, while the others were used to define asthma at 15–20 years. Asthma at that age was defined as positive responses to 'current wheezing' and 'asthma medications in the last year'.Two of the five STELAR cohorts (development data) were combined to develop predictive models using stepwise regression and frequentist, Bayesian and empirical Bayes regularization models. The remaining cohorts (validation data) were used to assess predictive performance using discrimination and accuracy measures. Analyses were performed in two populations - all children and a subgroup with reported wheezing between two and five years (high-risk population). Sex, eczema, sensitization to house dust mite and doctor's diagnosis of asthma in early childhood (4-7 years) were identified as asthma predictors at 15-20 years in both populations. Additional predictors in the general population included early wheezing symptoms and parental allergies, while in the high-risk population maternal allergies and pet in the house at one year were important for asthma prediction in adolescence. Sensitivity was higher in the general population, whereas positive predictive value was higher in the high-risk population. Although accuracy was good in both populations, the predictive ability of the models developed was limited.

Table of Contents

COPYRIGHT DECLARATION ..... **I**

ACKNOWLEDGMENTS ..... **II**

ABSTRACT ..... **III**

INDEX OF TABLES ..... **VIII**

INDEX OF FIGURES ..... **XI**

1 CHAPTER 1 – INTRODUCTION **1**

1.1 ASTHMA PHENOTYPES 4

1.1.1 Transient early-onset wheeze 6

1.1.2 Persistent wheeze 6

1.1.3 Late-onset wheeze 7

1.2 ASTHMA RISK FACTORS 8

1.2.1 Prenatal exposure 8

1.2.2 Atopy/Allergic sensitization 8

1.2.3 Sex differences 9

1.2.4 Hygiene hypothesis 9

1.2.5 Antibiotic use 9

1.2.6 Environmental risk factors 10

1.2.7 Genetic risk factors 10

1.3 ASTHMA PREDICTION MODELS 12

1.3.1 Asthma Predictive Index (API) and other models 12

1.3.2 Model generalizability 13

1.4 THEORETICAL BASIS OF PREDICTIVE MODELS 14

1.4.1 Development 14

1.4.1.1Scoring system based on univariable analysis 14

1.4.1.2Predictive models based on multivariable regression analysis 15
1.4.1.3Predictive models not based on multivariable regression analysis 16
1.4.2 Validation 17
1.5 AIMS AND THESIS OVERVIEW 18
2 CHAPTER 2 – VALIDATION OF CHILDHOOD ASTHMA PREDICTIVE MODELS: A SYSTEMATIC REVIEW 21
2.1 METHODS 23
2.1.1 Search Strategy 23
2.1.2 Data extraction 26
2.2 RESULTS 27
2.2.1 Target Population 27
2.2.2 Predictors 30
2.2.3 Outcome definition 32
2.2.4 Statistical modelling and predictive performance 35
2.2.5 Risk of bias assessment 39
2.3 DISCUSSION 39
2.4 CONCLUSION 42
3 CHAPTER 3 – ASTHMA IN ASHFORD: A 20-YEAR FOLLOW-UP 44
3.1 METHODS 46
3.1.1 The Ashford study 46
3.1.2 Statistical analysis 48
3.2 RESULTS 49
3.2.1 Differences between respondents and non-respondents 49
3.2.2 Definition of active asthma at age 20 years 51
3.2.3 Differences between asthmatics and non-asthmatics at the age of 20 years 51
3.2.4 Differences between paper and web questionnaire 55
3.3 DISCUSSION 56
4 CHAPTER 4 – DATA SOURCES 59
4.1 STELAR CONSORTIUM 60
4.1.1 Study populations 61
4.1.1.1ALSPAC 61
4.1.1.2Ashford 61

4.1.1.3Isle of Wight 62
4.1.1.4MAAS 63
4.1.1.5SEATON 64
4.1.2 Asthma e-Lab 66
4.2 DATA ANNOTATION 67
4.3 DATA AVAILABILITY AND HARMONISATION AMONG THE STELAR COHORTS 70
4.3.1 Identification of variables among the STELAR cohorts 70
4.3.2 Transformation of the STELAR variables 76
4.3.3 Matching of variables among the STELAR cohorts 78
4.4 CONCLUSION 78
5 CHAPTER 5 – COMPARISON OF STATISTICAL METHODS FOR THE PREDICTION OF ASTHMA AT 15–20 YEARS .. 80
5.1 METHODS 82
5.1.1 Definition of the target population 82
5.1.2 Definition of outcome 82
5.1.3 Potential predictive variables 83
5.2 STATISTICAL ANALYSIS 83
5.2.1 Model Development 84
5.2.1.1Classical stepwise regression 85
5.2.1.2Classical regularization regression analysis 87
5.2.1.3Bayesian regularization analysis 88
5.2.1.4Empirical Bayes regularization analysis 90
5.2.2 Model validation 92
5.2.3 Missing values in candidate predictors 95
5.3 RESULTS 97
5.3.1 Study population 97
5.3.2 Model development 100
5.3.3 External validation of model performance 102
5.4 DISCUSSION 103
5.4.1 Main findings 103
5.4.2 Comparison with existing asthma predictive models 105
5.4.3 Strengths and limitations of this work 107
5.5 CONCLUSION 108

6 CHAPTER 6 – FURTHER INVESTIGATIONS USING THE BAYESIAN MODEL 110

6.1 METHODS 111

6.1.1 Definition of target population and outcome 111

6.1.2 Potential predictive variables 112

6.1.3 Statistical analysis 112

6.2 RESULTS 114

6.2.1 Study population 114

6.2.2 Predictive models 117

6.3 DISCUSSION 123

6.3.1 Asthma predictors 123

6.3.2 Predictive performance variation 125

6.3.3 Conclusion and implications for future research 128

7 CHAPTER 7 – GENERAL DISCUSSION 129

7.1 MODEL VALIDATION IN EXTERNAL STUDIES AND CLINICAL SETTINGS 129

7.2 ASSESSMENT OF STATISTICAL METHODS 131

7.3 PREDICTIVE PERFORMANCE IN PRACTICE 131

7.4 OUTCOME AND PREDICTORS CHARACTERIZATION 132

7.5 ADDITIONAL CONSIDERATIONS 135

REFERENCES 137

APPENDIX A 148

APPENDIX B 159

APPENDIX C 174

APPENDIX D 176

Index of tables

Table 2.1: Characteristics of development studies included in the systematic review.....28
Table 2.2: Characteristics of validation studies included in the systematic review.....29
Table 2.3: Candidate predictors explored among development studies included in the systematic review.....32
Table 2.4: Outcome definition across studies.....34
Table 3.1: Characteristics of subjects at age 14 years by participation status in the 20-year follow-up.....50
Table 3.2: Results from the multiple logistic regression analysis on participation at age 20 years.....50
Table 3.3: Prevalence of features of asthma at the age of 20 years.....51
Table 3.4: Characteristics of the study population by presence of asthma at age 20 years.....52
Table 3.5: Respiratory and other symptoms by presence of asthma at age 20 years.....53
Table 3.6: Indoor and outdoor environmental risk factors by presence of asthma at age 20 years.....53
Table 3.7: Health and physical activity by presence of asthma at age 20 years.....54
Table 3.8: Results from the multiple logistic regression analysis on asthma at age 20 years.....55
Table 3.9: Characteristics of respondents at age 20 years by type of questionnaire.....56
Table 4.1: Characteristics of the STELAR cohorts.....65
Table 4.2: Metadata file created to annotate some of the variables collected using an Ashford questionnaire at age two years.....69
Table 4.3: Comparable variables among the STELAR cohorts downloaded from the e-Lab.....74
Table 5.1: Confusion matrix for binary classification.....93
Table 5.2: Discrimination measures obtained by confusion matrix.....93
Table 5.3: Definition of asthma at 15–20 years in children at high risk included in the development data.....98
Table 5.4: Definition of asthma at 15–20 years in children at high risk included in the validation data.....98
Table 5.5: Characteristics collected by age five years among the STELAR cohorts in high-risk population.99

Table 5.6: Analysis for asthma risk at 15–20 years in the high-risk population using a multiple imputation technique.....101
Table 5.7: Performance measures for the analysis in the high-risk population using a multiple imputation technique.....102
Table 6.1: Definition of asthma at age 15–20 years in the general population included in the development data.....114
Table 6.2: Definition of asthma at age 15–20 years in the general population included in the validation data.....114
Table 6.3: Characteristics collected by age 10 years among the STELAR cohorts in the general population.....115
Table 6.4: Characteristics collected by age 10 years among the STELAR cohorts in high-risk population.....116
Table 6.5: Analysis for asthma risk at age 15–20 years in the general population using multiple imputation technique.....119
Table 6.6: Analysis for asthma risk at 15–20 years in the high-risk population using a multiple imputation technique.....120
Table 6.7: Performance measures for the analysis in general population among models developed using a multiple imputation technique.....121
Table 6.8: Performance measures for the analysis in the high-risk population among models developed using a multiple imputation technique.....121
Table A.1: Odds Ratios estimated for predictors included into predictive models of development studies included into the systematic review.....155
Table A.2: Statistical methods used among studies.....156
Table A.3: Performance measures among studies and target populations.....157
Table A.4: Critical appraisal of the eight selected studies based on the CHARMS checklist.....158
Table C.1: Proportion of missing values among the STELAR cohorts and candidate predictors collected by age five years in the high-risk population.....174
Table C.2: Univariate analysis in the development data (ALSPAC and Ashford study) for the high-risk population.....175
Table D.1: Proportion of missing values among the STELAR cohorts and candidate predictors collected by age 10 years in the general population.....176
Table D.2: Proportion of missing values among the five STELAR cohorts and candidate predictors

collected by age 10 years in the high-risk population.....177
Table D.3: Univariate analysis in the development data (ALSPAC and Ashford study) for the general population.....178
Table D.4: Univariate analysis in the development data (ALSPAC and Ashford study) for the high-risk population.....179

Index of figures

Figure 1.1: Representation of normal and obstructed airways.....2
Figure 2.1: Flow diagram of the systematic review on asthma prediction.....25
Figure 2.2: Performance measures of API applied to children age five to eight years.....35
Figure 2.3: Performance measures of API applied to children age 10-11 years.....36
Figure 2.4: Performance measures of the PIAMA and Pescatore's risk score.....37
Figure 3.1: Children of the Ashford cohort.....43
Figure 4.1: Asthma e-Lab home page.....70
Figure 4.2: Results from searching on 'smoking' in the Ashford cohort.....71
Figure 4.3: Selection of variables of interest and inclusion in the online basket.....72
Figure 6.1: Definition of target populations in the STELAR consortium.....112
Figure A.1: Search on 14 July 2017 in Ovid MEDLINE(R) from 1946 to July week 2 of 2015.....148
Figure A.2: Search on 14 July 2017 in Ovid EMBASE from 1947 to 11 July 2017.....149
Figure A.3: The data extraction form used for the systematic review.....150
Figure B.1: The Ashford survey.....159
Figure B.2: Sample letter to possible participants.....173

Chapter 1 – Introduction

Asthma is a global health problem, affecting ~300 million people worldwide (1), a figure that is projected to increase to 400 million by 2025 (2). It is common in both western and developing countries (1, 2), and the World Health Organization suggests that it is the most common non-communicable disease among children. In the UK, an estimated 5.4 million people have asthma, of whom 1.1 million are children (3). Asthma adversely affects school and work performance (4), leading to a substantial burden on individuals and both direct and indirect costs for society. For example, the British annual healthcare costs associated with asthma are currently £1.1 billion, with at least £666 million spent on prescriptions. Additional costs include £160 million on general practice consultations, £143 million on disability claims and £137 million on hospital care (Asthma UK).

Asthma in children and adults includes a collection of clinical syndromes that to a greater or lesser extent limit expiratory airflow due to chronic inflammation, hyper-secretion of mucus and bronchial hyper-responsiveness (1, 3). During normal breathing, muscle bands surrounding the airways are relaxed, and air moves without obstruction. In people with airway obstruction, such as asthma, the airways walls are thickened and the bands of muscle surrounding them become tightened during asthma attacks (**Figure 1.1**). A person experiencing obstruction may feel short of breath, and the air moving out through the narrowed airways causes a whistling sound known as wheezing. Fortunately, this airway narrowing in asthma is generally reversible by using medications such as bronchodilators and corticosteroids.

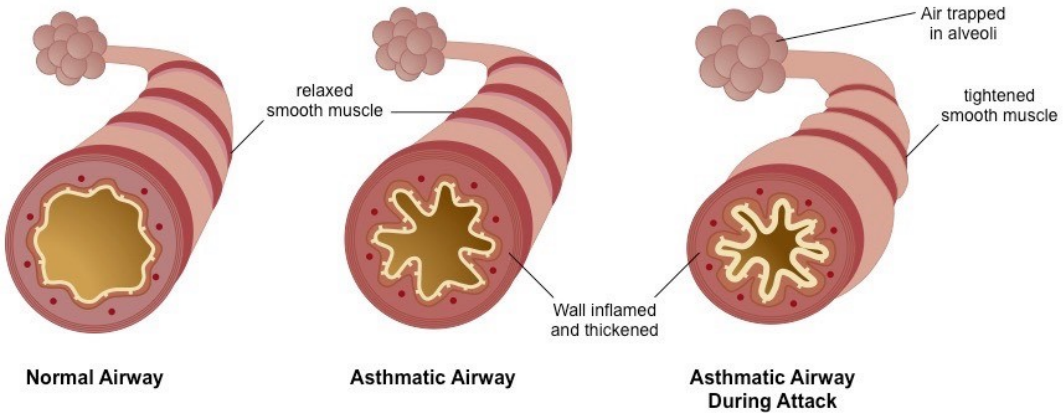


Figure 1.1: Representation of normal and obstructed airways (http://ib.bioninja.com.au/)

Thus, the most common asthma symptoms include wheezing, shortness of breath, chest tightness and cough (1, 3). These symptoms usually start during the first three years of life, with one in three children experiencing more than one episode of wheezing. They frequently abate during school age and sometimes recur in adolescence and adulthood (5). In established asthma, the most common triggers of symptoms are exposures to aero-allergens, such as those from pollen, house dust mite and animals; viral upper airway infections; exercise; cold air; exposure to tobacco smoke and ambient fine air pollution (1).

The diagnosis of asthma is based on a combination of clinical symptoms, measures of lung function and the exclusion of alternative diagnoses (1, 3). In children, it is useful to divide asthma by age at onset, specifically, preschool (age five years or younger) and school-aged (age over five years) asthma. In school-aged children and adults, a measurement of lung function is used to support the diagnosis of asthma and to evaluate the need for treatment adjustment (1, 3). However, in preschool children, asthma-like symptoms are common, and because a diagnosis cannot be supported by standard lung function tests (because of lung immaturity and the inability to perform accurate standard lung function measurements), clinicians must rely on symptoms such as coughing, breathlessness and wheeze (1, 3).

Specifically, the UK National Institute of Health and Care Excellence (NICE) has proposed an asthma diagnostic algorithm for adults and young people, based on four measures of lung function and inflammation, including spirometry (forced expiratory volume in 1s [FEV_1], forced vital capacity [FVC]

and the ratio of FEV_1/FVC); bronchodilator reversibility; fractional exhaled nitric oxide (FeNO) and peak flow variability (3). The algorithm starts with a spirometry test to identify airways obstruction; a bronchodilator reversibility test is then considered to confirm the obstruction and its reversibility (if any); the FeNO and peak flow variability tests are performed in cases of diagnostic uncertainty after spirometry. In practice, these objective examinations have not been integrated, and the accomplishment of lung function tests is generally difficult among preschool children, making the diagnosis problematic and unreliable until at least school age (6). When children are unable to perform objective tests for diagnosing asthma, physicians are advised to assess the condition based on symptoms and clinical judgment and review the child regularly until lung function measurements can be obtained.

In contrast, the Global Initiative for Asthma (GINA) guideline (1) has defined childhood asthma by recurrent patterns of a number of respiratory symptoms, such as wheezing, chest tightness and coughing, with these symptoms tending to be more prevalent during nights and early mornings, and by their response to asthma medications (bronchodilators and inhaled corticosteroids) (7, 8). However, the assessment of symptoms alone in clinical practice and epidemiologic studies may lead to misdiagnosis of asthma and inappropriate disease management. An incorrect diagnosis of asthma can occur due to misinterpretation of respiratory symptoms not related to asthma and/or the subjectivity of self-reported information within the interpretation of the term 'wheeze' (9).

Thus, the clinical definition of asthma in children is not entirely clear, and the relationship of early-life asthma to the common symptoms of wheeze, cough and breathlessness is not completely understood (10). Although these symptoms are the mainstay of measurement when attempting to diagnose asthma, their variation during the first five years of life makes it difficult to distinguish early-life asthma from factors that may increase the chance of developing the condition. To clarify the wheeze-asthma relationship in childhood, previous studies associated preschool asthma with factors relating to early symptoms, epigenetics and the environment (11, 12).

Asthma is thought to develop from an inappropriate immune response early in life caused by complex interactions between exposures and the host immune system. The heritability of asthma is high, and a large-scale, genome-wide study of asthma showed associations with genes that control

innate and adaptive immune components (13). There is a clear indication that the developing immune system is sensitive to environmental exposures (see section 1.2.7 Gene-environment interactions), suggesting the possibility of gene-environment interactions in its pathogenesis (14).

The variability in the expression of asthma-like symptoms observed in a clinical setting has prompted a move away from the concept that asthma is a single entity; asthma is now generally considered an umbrella term for several distinct conditions that share common clinical features (10).

1.1 Asthma phenotypes

A phenotype is defined as an observable property that arises from the interaction of genes and environmental exposures. Phenotypes are characteristics that can be observed and measured over a certain period of time (15-17). Asthma phenotypes have potential value for understanding and differentiating the mechanisms underlying the different conditions grouped under the term 'asthma' (16-18); these phenotypes could also improve our understanding of the natural history of asthma and could lead to more precise prevention strategies and to the development of effective personalised medicine (16, 17). Although the characteristics of childhood asthma overlap with those of several respiratory syndromes, including bronchitis, early-life exposures of children who exhibit transient symptoms appear to differ from those in whom asthma persists. The following three broad clinical wheeze phenotypes are recognised: atopic asthma, non-atopic asthma and virus-associated wheeze.

- Children with atopic asthma have cutaneous sensitisation to at least one allergen, with wheezing episodes that often occur in association with exercise or exposure to allergens. Early-onset atopic asthma is associated with severe wheeze—often persistent—and airway abnormalities (19).
- Children with non-atopic asthma have wheezing episodes without any atopic sensitisation; non-atopic asthma is often outgrown during childhood (19).

- Children with virus-associated wheeze have symptoms that occur only during an upper or lower respiratory tract infection. Pre-existing airway abnormalities are common (19), but a genetic or personal history of atopy is no more common than usual. Although this phenotype is usually outgrown by five years of age, there is concern that children with transient virus-associated wheeze might develop chronic obstructive pulmonary disease in adulthood (20).

The presence of wheeze is one of the key determinants of asthma. Wheeze can manifest in various ways and at different time points in a child's life. However, not all wheeze leads to asthma, particularly in children younger than age three years. Several studies have attempted to refine wheeze phenotypes in childhood; at the most general level, phenotypes have been characterised based on triggers or age at onset. In the past few decades, European Task Forces have proposed phenotypes of preschool children with wheeze, based on literature reviews and clinical knowledge (21, 22). Those phenotypes were based on periodicity or triggers, such as 'episodic viral wheezing' and 'multiple trigger wheeze' (21-23). In episodic viral wheezing, asthma-like symptoms are precipitated exclusively by viral colds; in multiple trigger wheeze they occur in the absence of colds but during exercise, laughing and other activities. Whereas it was assumed that episodic viral wheeze was associated with viral respiratory tract infections and that its prevalence decreased over time, multiple trigger wheeze was considered to be associated with chronic allergic airway inflammation and to precede asthma (21-23). Overall, a small number of studies reported on these trigger-related phenotypes but with conflicting results (23, 24). Thus, the ability of such phenotypes to distinguish which preschool children will or will not have asthma at school age remains limited, and they cannot distinguish, with sufficient clarity, different disease entities with separate aetiologies (16, 25).

Several longitudinal studies explored the progression of asthma symptoms from childhood to adulthood (9, 26). Temporal patterns of disease remission and relapse have been used to classify different wheezing phenotypes by age at onset during childhood (5, 27-29). The first study to explore these wheezing patterns was conducted in an unselected birth cohort based in Tucson, Arizona; Martinez *et al.* (5) used a subjective clinical approach to identify the following patterns of wheeze over time: never wheeze, transient early-onset wheeze, persistent wheeze and late-onset wheeze. Transient early-onset wheeze was defined as wheeze with onset before age three years and resolution by age six years; these children initially had poor lung function in infancy that later

improved but remained lower than that of the control group 'never wheeze', defined as children who had never wheezed by age six years.

1.1.1 Transient early-onset wheeze

In the Avon Longitudinal Study of Parents and Children (ALSPAC), Henderson *et al*. (27, 30) used a data-driven approach (longitudinal latent class model) based on repeated parental reports of wheeze from birth to seven years and identified an additional phenotype, prolonged early wheezing, which was characterised by wheezing between six and 54 months. Although this phenotype seemed to be a more severe form of transient wheeze identified in the Tucson study, early-life lung function was not available in the ALSPAC cohort thus hindering evaluation as to whether there is a distinction in premorbid lung function for this group of children compared to those with transient early wheeze. The results from ALSPAC have been replicated using similar methodology in other cohorts (31-33).

The Manchester Asthma and Allergy Study (MAAS) used a similar approach, integrating data from parental questionnaires and medical records to determine children's wheeze symptoms. Similar to the Tucson study, MAAS identified a transient early wheeze group characterised by wheeze between age one and five years. However, they found that lung function in these children remained impaired, compared to non-wheezing children, throughout childhood (31). Overall, risk factors for transient early wheeze include exposure to tobacco smoke, day care attendance, viral infections and family history of asthma (30, 34) (for more details, see section 1.2 Asthma risk factors).

1.1.2 Persistent wheeze

Children with persistent wheeze start wheezing early in life, and their symptoms continue into later childhood (5). One characteristic of persistent wheeze is diminished lung function by school age, with likely persistence into adulthood. Martinez *et al.* (5) found that, compared with the control group (non-wheezing), children with persistent wheeze initially had normal lung function, but their lung function worsened significantly by age six years. Using similar subjective methods to characterise children, groups from New Zealand (35) and Australia (36) found that children who were categorised as persistently wheezing continued to wheeze into adulthood and had consistently lower lung function. Data from primary care medical records in the MAAS cohort (31) allowed for stratification of persistent wheeze into two distinct subgroups: persistent controlled and persistent

troublesome wheeze. Children with troublesome wheeze were more likely to have a high symptom burden despite high doses of inhaled corticosteroids. In ALSPAC, those with persistent wheeze showed a similar pattern but with weaker atopy associations (27, 30). An intermediate-onset wheeze (onset between 18 and 42 months) was identified in ALSPAC; children in this category were more likely to have atopic sensitisation (particularly to house dust mite and cat), have poor lung function and subsequently be at risk of developing asthma in adolescence.

1.1.3 Late-onset wheeze

Late-onset wheeze is generally described as wheeze with onset after age three years that persists into later childhood. Atopic sensitisation is consistently associated with this phenotype of wheeze across different studies (30, 31). However, the association between late-onset wheezing, lung function and bronchial hyper-responsiveness differs among studies. For example, MAAS and ALSPAC found that late-onset wheeze was significantly associated with increased bronchial hyper-responsiveness and impaired lung function at age six years (30, 31), whereas the Prevention and Incidence of Asthma and Mite Allergy (PIAMA) (32) and Southampton Women's Study (SWS) (37) showed no such associations. Amongst environmental exposures, maternal smoking during pregnancy was a risk factor for late-onset wheezing in some (5) but not in all studies (31, 37).

By definition, these four temporal wheezing phenotypes (never wheeze, transient early-onset wheeze, persistent wheeze and late-onset wheeze) can only be identified retrospectively and thus cannot be used to establish immediate management and therapeutic strategies.

Other longitudinal studies investigated the higher prevalence of atopic disorders and characterised both the origin and evolution of childhood wheeze phenotypes (38, 39). These epidemiological studies suggested that respiratory morbidity, in childhood and adulthood, was associated with early onset of symptoms, early impairment of lung function and genetic disposition, leading to the idea that a combination of genetic and environmental factors might predispose to wheezing disorders as a consequence of altered immune or airway development. However, genetic changes at a population level cannot explain the rapid increase in asthma prevalence in the last few decades; early environmental exposures might be key factors in the inception of paediatric asthma and wheeze.

1.2 Asthma risk factors

1.2.1 Prenatal exposure

The development of asthma is thought to begin perinatally, that is, *in utero* (40-42), and may become problematic in childhood and later life. Indeed, prematurity and a suboptimal intrauterine environment may lead to lung impairment and can affect later respiratory health. Prenatal factors that have been shown to be responsible for the development of asthma include maternal dietary intake (43), maternal body composition and maternal endocrine status (44). There is also evidence that transgenerational factors might disrupt the long-term health of offspring. A well-recognised transgenerational factor is maternal birthweight, which influences the wellness of the foetus and child during early life (45). Foetal growth, which is linked to childhood respiratory function, depends on the uptake of nutrients from a complex maternal supply line: nutrients supplied to the foetus reflect placental function as well as maternal physiological adaptations to pregnancy, maternal intake and absorption of nutrients and foetal nutritional demand (43, 46). Although the link between foetal and maternal nutrition is indirect, the idea that maternal diet might influence her newborn's respiratory and immune development has received support from animal studies (47). The influence of breastfeeding on the risk of childhood asthma and atopy is controversial. Some studies have shown breastfeeding to be a protective factor (48), while others reported higher rates of allergy and asthma among breastfed children (49, 50). Thus, it is likely that asthma and atopy are influenced not by the act of breastfeeding itself, but by maternal exposures through the milk.

1.2.2 Atopy/Allergic sensitization

Another critical risk factor for asthma development is allergic sensitisation determined by the immune system, which is vulnerable to perinatal insults in a manner similar to that of airway development. The perinatal immune system is finely balanced, to ensure an appropriate response to an invading pathogen and to minimise risk of injury to the host (51). The deviation of the immune system toward a particular cytokine response (such as T-helper type 2 [Th2] cells), is often associated with atopic asthma (52). Infants and children who fail to convert from a foetal Th2 to a mature T-helper type 1 (Th1) immune phenotype are more likely to be sensitised and develop atopy (53).

However, this failure of the immune system response occurs mainly in individuals who are exposed to more detrimental environmental risk factors and/or are more genetically susceptible.

1.2.3 Sex differences

Sex differences are known to be associated with the development and persistence of asthma in a time-dependent manner. The incidence and prevalence of asthma are higher among boys than among girls until age 13–14 years (54), and a greater proportion of adolescent males have remission of asthma (55, 56). Boys have more severe asthma before age 12 than girls (57); conversely, adult females have more severe asthma than males (55). However, several studies through puberty have shown a greater incidence of asthma among adolescent males, with more hospital admissions and longer hospital stays (58-60). Perhaps this finding is related to sex differences in immune function as well as hormone levels across the lifespan (61, 62).

1.2.4 Hygiene hypothesis

Family size and the number and order of siblings are associated with the risk of developing asthma. The hygiene hypothesis suggests that infants who are exposed to many infections, including those caused by bacteria and parasites, might be protected from asthma development in childhood (63). Although this theory has been supported by some studies of allergy prevalence (64), it has been partially refuted by recent studies (65, 66). Children of parents with lower socioeconomic status tend to have greater morbidity from asthma (67), although the findings are mixed (68).

1.2.5 Antibiotic use

The use of antibiotics has been associated with early wheezing and asthma (69, 70). Frequent antibiotic use might represent a surrogate marker for a higher number of infections in early life. Infections of the lower respiratory tract affect early childhood wheezing, but it is unclear whether infection alone has a role in the development of persistent asthma. For example, infection of the lower respiratory tract in genetically susceptible infants who are sensitised to inhalant allergens, such as house dust mite, cat and mixed grass, may lead to deviation toward an immune response which promotes asthma.

1.2.6 Environmental risk factors

Maternal smoking during pregnancy affects the responsiveness of the foetal immune system, via cord blood mononuclear cells, to allergens (71) and contributes to limited airway function in newborn infants as well as poor respiratory health during childhood. Exposure to environmental tobacco smoke during early childhood has also been associated with impaired lung function (45). Other pollutants that can harm neonatal and childhood respiratory outcomes include indoor environmental factors, such as mould and harmful emissions from household cleaning agents and paint, and outdoor environmental factors, such as particulate matter pollution, sulphur dioxide, nitrogen oxides, ozone, extreme temperatures and high humidity (45).

Together these findings suggest that aspects of the *in utero* and early-life environments are significant determinants of later health and that factors influencing foetal growth, as well as the environment where the child lives, might also affect respiratory and immune development.

1.2.7 Genetic risk factors

Approaches to the discovery of genes related to asthma have ranged from candidate gene association studies and genome-wide linkage studies, to more recent genome-wide association studies (GWAS). The results of these studies have highlighted the importance of genetic variants in or near genes that are implicated in the development of asthma or allergic diseases (72, 73). Large studies, including the European GABRIEL Consortium, a multidisciplinary study to identify the genetic and environmental causes of asthma in the European community (13), and the US EVE Consortium (74), identified a number of genetic variants associated with asthma in children, in particular in the chromosome region *17q21*. However, the predictive value for this chromosome region is low, with specificity being 75% and sensitivity only 35%. Other genes, such as *IL33*, *IL1RL1* and *IL18R1*, have also been implicated in asthma development (75). A different approach using transcriptomic analysis of peripheral blood in children with virally induced wheeze and exacerbations found that there are distinct microRNA profiles in CD8 and T cells, with reduced regulation of microRNA-146a/b and microRNA-28-5p (76).

The results of genetic association studies have not been consistent across different populations, suggesting that genes that confer susceptibility to asthma might cause disease only in the presence of specific environmental exposures, that is, through gene-environment interactions (77). An example of how genetic susceptibility can be modified by environmental exposure is found in the interaction between the glutathione S-transferase M1 genotype and ozone exposure. Asthmatic children with a homozygous deletion of the glutathione S-transferase M1 gene appear more susceptible to the deleterious effects of ozone on small airway function and derive greater benefit from antioxidant supplementation (78).

Environmental factors can also interact with genetic risk by influencing the transcription of asthma susceptibility genes. For example, *in utero* exposure determined by maternal nutrition, smoking, body composition and maternal endocrine status could influence foetal physiological development directly by modifying gene expression.

Environmental factors can also interact with the genome via epigenetic mechanisms, which are heritable modification of gene expression by mechanisms that do not change the DNA sequence (79). Epigenetic modification in early life allow environmental factors to modify the expression of a child's genome with the potential to cause asthma and/or atopy. Several mechanisms of epigenetic modification of gene expression have been discovered (80). First, gene-regulatory regions can be methylated or demethylated; this process alters gene expression in a graded fashion. Second, the structure of the protein-DNA complex can be modified by histone chromatin acetylation. There is widespread demethylation and methylation during gametogenesis and during early embryogenesis just prior to blastocyst development; any modifications that occur then are stable throughout subsequent somatic differentiation. Epigenetic mechanisms have been shown to influence expression of transcription factors that control T cell lineage (81). A recent study, for example, showed that, in a mouse model *in utero*, dietary methyl donors increased the severity of allergic airway disease through epigenetic mechanisms (82). Inclusion of genetic factors in prediction models might increase their predictive performance and improve forecasting of asthma development. However, this information is not readily available in clinical practice and its inclusion would limit the routine use of a predictive model.

1.3 Asthma prediction models

It has proved difficult to predict which children with wheeze will have persistent symptoms of asthma later in life based on temporal patterns, environmental exposures and other risk factors. Being able to predict an individual's risk of developing asthma in adolescence and young adulthood would enable physicians to identify high-risk children and to forecast the development, progression or remission of the condition. It would also be helpful to parents, and it would assist in the design of, and recruitment to, studies of novel disease-modifying treatments. On the other hand, accurately identifying children at low risk of asthma would help prevent over-treatment of children with wheezing symptoms, thereby decreasing the incidence of side effects (83) and reducing the cost of healthcare services (84).

1.3.1 Asthma Predictive Index (API) and other models

In studies of asthma prediction, several risk profile models have been proposed to identify preschool children who are at high risk of asthma at school age (85, 86). The best known of these, the Asthma Predictive Index (API), is based on the Tucson Children's Respiratory Study (87). A positive API score requires recurrent wheezing episodes during the first three years of life (less than three wheezing episodes per year for the loose API criteria and three or more wheezing episodes per year for the stringent API criteria) and one of two major criteria (doctor's diagnosis of eczema or parental asthma) or two of three minor criteria (doctor's diagnosis of allergic rhinitis, wheezing without colds or eosinophil count ≥ 4%). A positive stringent API score by age three years was associated with a 77% chance of active asthma at age 6–13 years; children with a negative API score at age three years had less than a 3% chance of having active asthma during school age.

Several other predictive models have subsequently been proposed, such as the modified API (mAPI) and the Prevention and Incidence of Asthma and Mite Allergy (PIAMA) risk score (88), but they have shown limited clinical relevance, and their applicability in other populations with different clinical as well as sociodemographic characteristics is uncertain (89-92). Modifications of the initial API, which resulted in the mAPI (93) and the University of Cincinnati Asthma Predictive Index (ucAPI) (94), included parental history of physician-diagnosed asthma as a major criterion by which to identify children at high risk of asthma development.

The PIAMA study, which used data from 2,171 children with wheeze in the Netherlands, showed that male sex, post-term delivery, parental education, parental inhaled medication, wheezing frequency, wheezing/dyspnoea apart from colds, respiratory tract infections and eczema were predictors for asthma at age seven and eight years. In 2003, Kurukulaaratchy *et al.* (95) used data from 1,456 children in the Isle of Wight birth cohort to develop a predictive model based on family history of asthma, recurrent chest infections in the second year of life, atopic sensitisation at four years of age and absence of recurrent nasal symptoms in the first year of life. Those factors were associated with an increased risk of wheezing episodes at age 10 years. Matricardi *et al.* (96), in the German birth cohort Multicentre Allergy Study (MAS), which followed 441 children with wheezing before age three years, reported that parental atopy (including asthma, hay fever and atopic dermatitis) was an independent predictor of their child's wheezing at age 11–13 years. Vial Dupuy *et al.* (97), in a cohort of 200 infants less than two years of age with recurrent wheezing, reported that family history of asthma, eczema and sensitisation to common food and/or aeroallergens were predictors of persistent asthma at age six years. van der Mark *et al.* (98), in 438 children aged one to five years who attended primary care for recurrent coughing, wheezing and shortness of breath, identified family history of asthma, allergy and sensitisation to aeroallergens as predictors of asthma at age six years.

1.3.2 Model generalizability

Those predictive models differ mainly in terms of candidate predictors tested, target populations and age at outcome assessment, all of which affect the predictive performance and make comparisons between them difficult. The differences between them call into question whether they can be used successfully in other populations and emphasise the need to test them in different populations to give clinicians the confidence that they are both generalisable and clinically useful for prediction. These predictive models would need to be tested not only in populations different from those used to develop them, but also against each other in the same populations.

Most existing asthma predictive models, including the ones discussed above, were developed using frequentist statistical methods and were internally validated; that is, the same datasets were used both to develop the predictive model and to assess its performance, leading to potentially biased predictions and limiting the generalisability of the results in other populations (99, 100). Application of advanced statistical techniques to develop predictive models and use of large longitudinal samples have been limited in this field and may improve the accuracy as well as performance of asthma prediction models.

1.4 Theoretical basis of predictive models

Predictive models are statistical methods used in different fields, such as finance, education, medicine and law, to assist a decision-making process by predicting future outcomes (events) based on the characteristics of the data (101). Once a model is developed, it should be validated using independent data to evaluate its predictive accuracy. In medicine, predictive models can be used to predict the probability that a subject with a given set of predictors will develop a particular health outcome. To be useful in clinical practice, such models should predict the clinical outcome accurately and be relatively easy to use (102). The analytic process for developing clinical prediction models can be summarised by two main stages: 1) model development and 2) model validation.

1.4.1 Development

1.4.1.1 Scoring system based on univariable analysis

One simple method for developing a predictive model uses the results that are derived from a univariable analysis. For this method, associations between candidate predictors and a clinical outcome are determined by univariable statistical tests, such as the Student *t*-test, chi-square test or Fisher's exact test. Predictors that are associated significantly with the outcome are included in the final predictive model. Examples of this method are the API index (87) and mAPI (93).

Although this analysis can be performed with statistical techniques that are comprehensible by clinicians and accessible using basic statistical packages, this method has one main limitation. When only univariable analysis is performed, factors that are not associated independently with the

outcome and possibly related to one another can still be included in the final model, thereby reducing predictive accuracy (103).

1.4.1.2 Predictive models based on multivariable regression analysis

Multivariable analysis within a classical or Bayesian framework can be used to develop a predictive model. Linear, logistic or Cox regression techniques can be applied depending on the nature of the outcome (e.g., continuous or binary variable) and the aim of the research (102, 104, 105). This type of analysis resolves the limitation that are encountered when univariable analysis is used as the basis of a predictive model. Multivariable techniques allow for inclusion of a number of candidate predictors in the model leading to an improved assessment of the association of each independent candidate predictor with the clinical outcome. Most predictive models have been developed using this technique (88, 95, 96, 106-108).

In the case of binary outcomes, where logistic regression is used, the relationship between a predictive factor and the outcome is expressed by an odds ratio, a measure of effect that is used frequently in the medical literature and is relatively easy to interpret. By using a simple logistic equation, this method calculates the individualised probability that each subject will experience the outcome of interest. Inclusion of predictors that do not improve the predictive performance makes the model overly complicated and difficult to apply in clinical practice; hence, unnecessary variables should be removed for a more efficient predictive model. Deciding which predictor(s) to remove can be done by using clinical knowledge or statistical methods. Common statistical methods to perform variable selection include forward, backward and stepwise regression (102, 105). Although those methods are widely used, they have many disadvantages. For example, when the number of subjects is small and the outcome prevalence is low compared to the number of predictors included, the selection is unstable, the estimates of regression coefficients can be extreme and the performance of the predictive model tends to be overestimated. To overcome those issues, regularisation methods have been suggested, as they perform variable selection and estimate regression coefficients simultaneously, thereby improving accuracy and predictive performance (105, 107-110). Details on some of these methods are reported in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years, paragraph 5.2.1).

Although multivariable analysis based on regression is a powerful and widely used tool to develop prediction models, it is not without disadvantages. Specifically, independent variables may interact with one another. Although an investigator can search for such interactions, if many independent variables are being evaluated, an evaluation of all two-way interactions is at the very least tedious, and a comprehensive evaluation of higher-order interactions is all but impossible. Also, in using logistic regression, assumptions are made about the linear relationships between the dependent and independent variables, where in some cases this type of relationship does not exist. Researchers can investigate whether nonlinear relationships are present and, if so, whether transformations can resolve the nonlinearities. This work is time-consuming, and there may be instances in which complex nonlinear relationships cannot be modeled appropriately. In these instances, other methods that are more capable of assessing nonlinear relationships among variables may provide more accurate predictions.

1.4.1.3 Predictive models not based on multivariable regression analysis

Predictive models that do not use multivariable regression analysis, such as Machine Learning algorithms, are gaining in popularity as an alternative approach for prediction and classification problems. These methods include classification trees, random forests, artificial neural networks, support vector machines, and other algorithms. These methods have been used sporadically for clinical prediction for some time (111), but the growing availability of increasingly large, voluminous, and rich databases have increased interest in exploiting them(112). Notably, these methods have not been used widely in asthma prediction.

Machine Learning methods overcome some limitations of regression analysis as they can account for nonlinear relationships of independent variables with one another and with the dependent variable. These methods have this capability because they do not rely on the assumption of linear relationships that are inherent in regression analyses. Correspondingly, they free investigators from the tedious and sometimes unsuccessful search for these relationships and theoretically optimize the predictive accuracy of models when such relationships are present. However, this theoretic advantage, which allows one to determine otherwise indiscernible data patterns, is a prominent disadvantage for clinical translation. Although logistic regression equations and related scoring

systems can sometimes be tedious to use, they nevertheless can be written explicitly and disseminated easily.

1.4.2 Validation

The predictive performance of a model is evaluated using independent data that have structural and population characteristics similar to those of the data from which the model was developed (99, 105, 110). Predictive performance can be evaluated using internal or external validation. A common internal validation approach consists of splitting a dataset into two parts: one part is used to develop the model and the other to measure its performance (110). Other internal validation methods have been developed. A variation of the split-sample method is the *k*-fold cross-validation approach. The data are partitioned into *k* equally sized subsets (109, 113). First, *k-1* subsets are used to develop the predictive model, and the remaining subset is used to evaluate its predictive ability. This approach is repeated *k* times, and results are averaged to produce individual estimates of regression coefficients and standard errors. Further details on validation techniques are reported in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years).

Whereas internal validation addresses the stability as well as replicability of the selected predictors and the results, external validation refers to generalisability of findings to similar populations. External validation is considered a stronger assessment of a predictive model than internal validation and requires an assessment of the applicability of a predictive model in independent samples with subjects from different but similar populations that were not used to develop the model.

There are several standard performance measures by which to evaluate and capture different aspects of predictive models. They can be distinguished into calibration and discrimination measures (102, 105, 114). Calibration measures refer to the agreement between observed and predicted values; commonly used methods include the calibration plot and the Hosmer-Lemeshow goodness-of-fit test. Discrimination measures refer to the ability of a model to distinguish between subjects with and without the outcome of interest. The most popular discriminative measures are sensitivity, specificity, positive and negative predictive values and the area under the receiver operating characteristic (ROC) curve (105, 114, 115) additional details on performance measures and their

interpretation are reported in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years).

1.5 Aims and thesis overview

Current asthma prediction models were developed to characterize preschool children who are at risk of asthma at school age, while there are no models designed to predict asthma in adolescence and young adulthood. In addition, the majority of asthma predictive models used in early childhood were developed using frequentist statistical methods and internal validation techniques, leading to poor generalisability of results in external populations. As a result the aims of this thesis are as follows:

1. to review externally validated asthma predictive models together with the original development studies and explore their predictors, definitions of target outcome, statistical methods and predictive abilities (Chapter 2);
2. to describe the Ashford birth cohort and to design a paper/online survey to collect information at 20 years of follow-up (Chapter 3);
3. to annotate the Ashford information collected over 14 years and to harmonise information from five asthma UK cohorts (Chapter 4);
4. to develop a statistical model that predicts asthma between ages 15 and 20 years by using information gathered in early childhood (birth to five years) and to investigate several statistical techniques, including stepwise logistic regression, two regularisation logistic models developed in a frequentist framework, least absolute shrinkage and selection operator (LASSO) and elastic net (EN), as well as empirical Bayes and Bayesian regularization methods, in order to determine which statistical method provides the best predictive performance (Chapter 5);
5. to explore whether prediction of asthma between 15 and 20 years can be improved by the inclusion of additional information collected in mid-childhood (five to 10 years) (Chapter 6).

The layout of the chapters of the thesis is as follows. In Chapter 2, I systematically review studies that undertook external validation of existing predictive models of asthma in later childhood and adolescence, alongside the studies in which the models were originally developed. In Chapter 3, I

introduce the Asthma in Ashford birth cohort study and describe the methods used to develop an online and paper survey for its latest follow-up, in which participants, now aged 20 years, completed a questionnaire detailing asthma-related symptoms and risk factors. Findings, including comparison of respondents and non-respondents, children with and without asthma and response rates from paper- and web-based questionnaires are presented.

In 2013, the Ashford cohort joined the Study Team for Early Life Asthma Research (STELAR) consortium (116) together with four other UK asthma birth cohorts based in Bristol, Isle of Wight, Manchester and Aberdeen. In Chapter 4, the harmonisation procedure among the five STELAR cohorts is described, and details about the coding of variables and their definitions are provided. Information collected at comparable ages over 10 years across the UK cohorts is harmonised to define candidate predictors tested in my predictive models, and data from the Ashford cohort of study subjects now aged 20 years and data collected by the other STELAR cohorts at a comparable age are used to define asthma in adolescence. In Chapter 5, I explain the process of developing predictive models by using the statistical techniques mentioned above, along with external validation procedures, and I identify which statistical method works best to identify predictors of asthma between 15 and 20 years. For this purpose, only information collected by age five years is included in the predictive models, and the harmonised STELAR data are split into two parts; the ALSPAC and Ashford studies are used to develop the predictive models, and the remaining three cohorts are used to externally assess the predictive ability of the models developed. The analysis presented in Chapter 5 was performed in a subgroup of STELAR children ('high-risk population') whose parents reported that they had wheezing episode(s) between the ages of two and five years. In Chapter 6, I use only the best statistical method identified in Chapter 5 to examine whether prediction can be improved by adding extra information collected from ages five to 10 years, including respiratory symptoms and certain clinical measures. The analyses presented in Chapter 6 were performed not only in the high-risk population defined in Chapter 5 (i.e. children with wheezing symptoms at preschool age) but also in the full STELAR population, which includes all children recruited at birth among the cohorts (i.e. 'general population'). Finally, in Chapter 7, the clinical implications of my findings, strengths and limitations of the thesis, along with opportunities for future studies are discussed.

Chapter 2 – Validation of childhood asthma predictive models: a systematic review

Wheeze, cough, shortness of breath and chest tightness are cardinal symptoms of asthma. Wheeze tends to start in early life, often disappears during school age, and occasionally recurs later in life; in some children symptoms persist throughout childhood into early adulthood (1, 117). In clinical practice, accurate models for the early identification of children at high risk of late asthma would facilitate patient-physician communication based on information that is more objective. Physicians would be able to inform individuals of their risk of developing asthma later in life and to notify patients and their caregivers whether further testing and/or treatment are advisable (102, 118-120). The early exclusion of risk of asthma later in life could help prevent unnecessary tests and over-treatment and could also reduce costs (83, 84).

Models that predict disease are clinically valuable when they accurately assess the risk of developing the condition in the future and when they have an impact on clinical decision-making, patient outcomes and health costs (102, 104, 119-124). In medical research, predictive models can be used to screen and stratify patients for experimental studies, based on their risk of developing the health outcome of interest (104, 118, 121). Such models have clinical utility not only when they can distinguish patients with and without the disease with high accuracy, but also when their performance is reproducible across different populations and clinical settings (104, 121, 122, 125).

It is necessary to test and validate predictive models in external studies, using different populations with comparable characteristics and similar clinical settings, as this enables researchers to quantify their predictive ability and assess their generalizability (99, 104, 119, 126-128). The appropriate specification of the target population for external validation is crucial for assessing the validity and clinical utility of predictive models. The importance of validation has been demonstrated many times in different fields of medical research (129, 130), and significant reductions in the predictive ability of a model when applied to a new dataset have been reported consistently (130). Despite this, only a few asthma studies validated their models in different populations and settings; methods for internal, or apparent, validation are commonly used, but such approaches can lead to over-optimistic results (99, 104, 110, 122).

Many prediction models have been proposed to forecast the development of asthma in children in the general population (87, 107, 108, 131, 132) and in high-risk groups (88, 89, 94, 98, 106, 133, 134), but their predictive performance, clinical usefulness and reproducibility in external populations are limited, making them difficult to use in clinical practice and/or research (85, 86). Despite the lack of validation and uncertainty of predictive performance in different populations, the U.S. National Asthma Education and Prevention Program Expert Panel Report 3 (EPR-3) recommends the use of the modified asthma predictive index (mAPI) (135). In contrast, two recent comprehensive systematic reviews (85, 86) concluded that none of the existing models can robustly predict or rule out asthma. Those reviews focused on model development studies, without exploring the reliability and usefulness of the predictive models in different populations and clinical settings. In the current study, we focused on validation studies together with the original development studies, and we assessed their statistical methods, predictive abilities and clinical usefulness in children from both general and high-risk populations. This work has been published in *Clinical and Experimental Allergy Journal* (CEA) (136).

2.1 Methods

2.1.1 Search Strategy

In this review, I focus only on studies that externally validated existing asthma predictive models, alongside the studies in which they were originally developed. My methods were based on the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (137) and the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statements (138). The protocol was registered on PROSPERO (CRD42016035727), an international database of prospectively registered systematic reviews in health and social care. Registration enables investigators to identify ongoing and unpublished reviews that address their topic of interest; this helps to avoid unintended duplication. The main features from the review protocol were recorded and approved as a permanent record.

On 11 July 2017, I performed an electronic search of two databases, MEDLINE and EMBASE, using both free text and MESH terms. The search strategy was established initially for the MEDLINE database (Ovid interface) based on papers published from 1946 to the first week of July 2017. First, a list of words related to asthma was defined, and appropriate truncation and wildcards were used to find all related terms. The spelling of all the search terms and the number of records identified were checked to monitor the quality and validity of the search. Then, the keyword 'asthma' was adopted to map the related MESH terms. Finally, keywords and MESH terms were combined by using the Boolean operator 'OR'. A similar approach was used for the predictive part; several keywords with wildcards were used to identify relevant records that contained terms such as 'predictive score', 'predictive model' and 'predictive performance'. MESH terms for this part of the search were not available. The two blocks of the search were combined by an 'AND' logical operator, and then limitations to humans, children (0 to 18 years) and English language were applied. The search strategy from MEDLINE was then adapted for the EMBASE database, and the search findings from the two databases were integrated. Details of the search strategy for both MEDLINE and EMBASE are reported in **Figure A.1** and **Figure A.1**.

In addition, further studies were traced through cross-checking of reference lists from identified relevant papers. Studies of all designs were included only when they (i) investigated the prediction

of asthma in external validation studies and in the original studies that had developed the validated models; (ii) performed outcome assessment in children between the ages of 6 and 18; (iii) tested at least three candidate predictors in the model development; (iv) evaluated at least one performance measure, including overall accuracy, discrimination and calibration measures; (v) were written in English.

I excluded non-original studies, such as letters, editorials and conference papers. Overall, more than 900 papers were identified as potentially eligible; they were screened for eligibility based on title, abstract and full text when necessary, by two independent reviewers (SC and DM). The reason for exclusion was recorded. A flow diagram of the selection process is shown in **Figure 2.1**; it describes the selection of studies, including the number of records identified from different sources, the number of studies excluded from the initial screening and the reasons for exclusion.

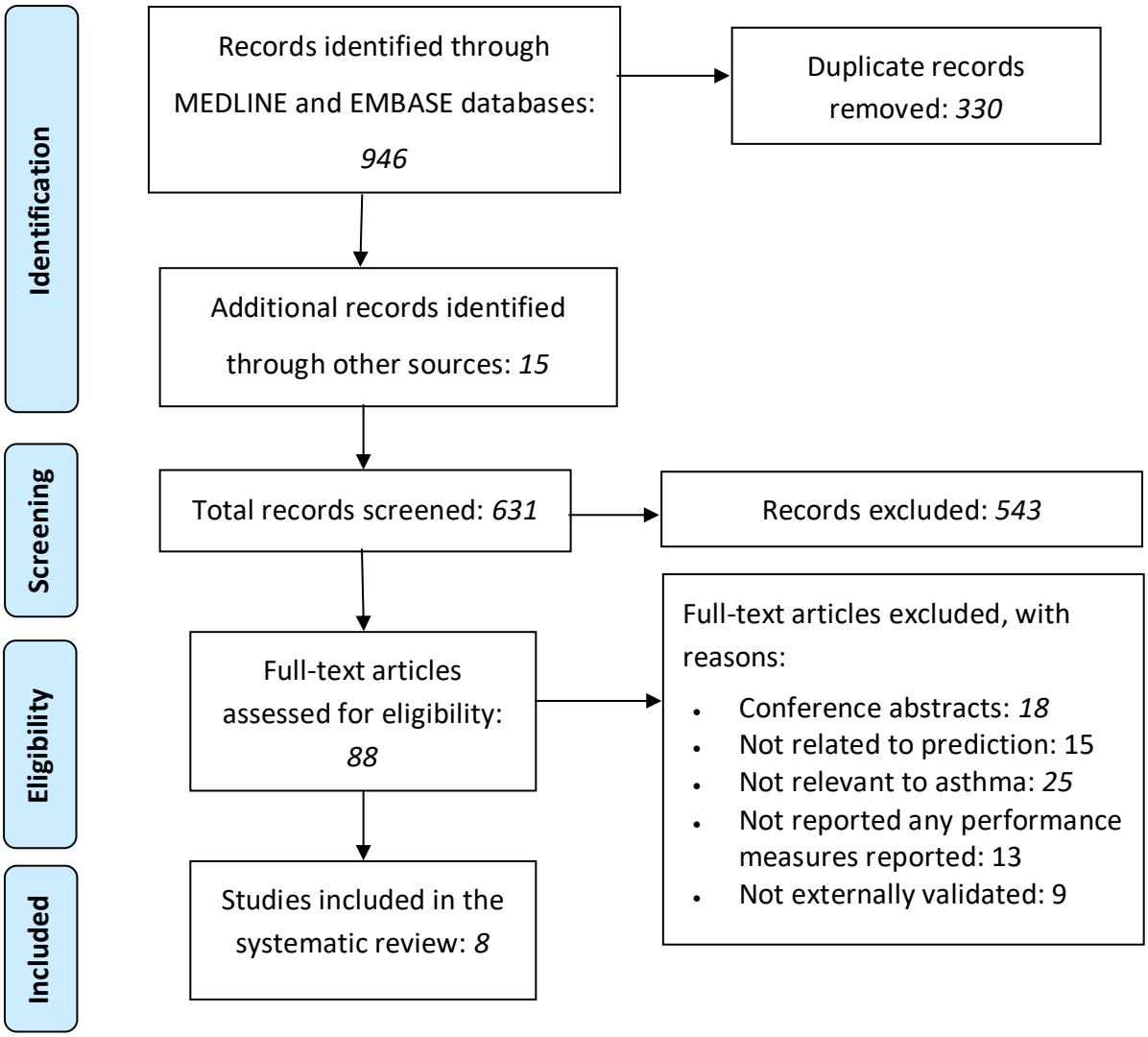


Figure 2.1: Flow diagram of the systematic review on asthma prediction

2.1.2 Data extraction

Data extraction was performed in duplicate by two reviewers (SC and DM) using a form (**Figure A.3**) developed in accordance with the CHARMS checklist (137). A wide range of information was extracted from each study in order to assess the validity of the study and the accuracy of the predictive model. This information included details of the target study population (such as country and ethnicity); study design; sample size at recruitment and year of recruitment; participants' eligibility and enrolment method; participants' characteristics at baseline and follow-up; definition of asthma; count of asthma events at each assessment to enable comparison across studies and to evaluate the risk of 'overfitting', which occurs when the number of predictors is too high relative to the sample size; number and description of candidate predictors, with details of their definitions and the way they are modelled in the analysis; amount of missing data reported in both outcomes and predictors and method(s) used to handle them; selection method for identifying final predictors and type of model used for disease prediction; model accuracy at each asthma assessment and performance indicators, such as sensitivity, specificity, calibration and classification measurements; methods used for internal and/or external validation; interpretation of findings, discussion of assumptions underlying the statistical methods used and comparison with findings and methods in previous studies. The reviewers resolved any discrepancies through discussion with the other two authors (CM and PC) until consensus was reached. I extracted performance measures and calculated predictive values and likelihood ratios, if they were not reported.

2.2 Results

2.2.1 Target Population

Studies that aimed to identify children who will have asthma in later life could be grouped into the following two categories: a) studies of the general population [with two studies recruiting participants at birth (87, 108) and four studying generally healthy children recruited through healthcare practices, schools or local registries (91, 107, 131, 139)]; and b) studies of children at high risk of late asthma [with two studies recruiting infants born to allergic parents (89, 94), eight studies of symptomatic children recruited at healthcare practices or schools (88, 90-92, 98, 106, 133, 140) and a single study looking at children attending hospital with asthma symptoms] (134).

In total, only three prediction models were externally validated and therefore considered in this review: 1) the API (87), which was validated in two studies (90, 139); 2) the Prevention and Incidence of Asthma and Mite Allergy (PIAMA) tool (88), which was validated in one study (91), with another study assessing the predictive capabilities of both API and PIAMA (92); and 3) a simple asthma prediction model developed by Pescatore *et al.* (108), which was validated in one study (140). We reviewed these three development studies (87, 88, 108) (**Table 2.1**) and the five validation studies (90-92, 139, 140) (**Table 2.2**). The API was the only tool developed in a general population, but the validation of the main characteristics of those studies occurred in populations with widely different sample sizes, from a minimum of 130 participants (92) to more than 3,000 children (139). Study populations included cohorts from the UK (108, 139), central Europe and Scandinavia (88, 90, 91, 140), USA and South America (92). The prevalence of asthma at outcome assessment varied from 5.8% (91) to 66.7% (92).

stics of development studies included in the systematic review

	Country	Study name (recruitment year)	Inclusion criteria	Sample size (% of those eligible)	Prediction rule	Age at outcome assessment (years)	Asthma prevalence at outcome n (%)
- *uez*	USA	The Tucson Children's Respiratory Study (TCRS) (1980-1984)	Healthy new-born babies recruited in primary healthcare practices	1,246 (78)	Stringent API and Loose API	6 8 11 13 6-13	112 (11) 113 (14) 150 (16) 122 (17.5) 250 (35)
i	The Netherlands	PIAMA (1996-1997)	Children with asthma-like symptoms at age 1-4 years	2,171 (55)	PIAMA	7-8	362 (16.7)
tore	UK	Leicestershire Respiratory Cohort Studies (1993-1997)	Children aged 1 to 3 years with parent-reported wheeze or chronic cough	2,444 (36)	Risk score	6-8	345 (28.1)

Table 2.2: Characteristics of validation studies included in the systematic review

Target population	ID	Country	Study name (recruitment year)	Inclusion criteria	Sample size (% of those eligible)	Prediction rule	Age at outcome assessment (years)	Asthma prevalence at outcome n (%)
General population	*Leonardi 2011*	UK	Leicester Respiratory Cohort (1998)	Children, between 0-4 years, registered in local child health database	3392 (79)	Stringent API and Loose API	7	220 (11.5)
							10	147 (10.5)
High-risk population	*Devulapalli 2008*	Norway	Environment and Childhood Asthma study (NA)	Bronchial obstruction by age 2	612 (NR)	Stringent API and Loose API	10	239 (52.07)
	Rodriguez-Martinez 2011	Colombia	NR (2006-2007)	Parent-reported wheezing	130 (NR)	Stringent API and Loose API	5-6	33 (35.5)
						PIAMA		66 (53.6)
	Hafkamp-de Groen 2013	The Netherlands	Generation R (2002-2006)	≥1 positive response to wheeze at 1 year	2877 (73)	PIAMA	6	168 (5.8)
	Grabenhenrich 2014	Germany	The German Multicentre Allergy Study (MAS-90) (1990)	Children with wheezing or cough in in the previous 12 months at the age of 3years	140 (17)	Risk score	8	28 (20)

Abbreviation - NR: not reported.

2.2.2 Predictors

Castro-Rodriguez *et al.* (87) used information from two- and three year old children to develop two prediction models (the 'loose' and 'stringent' versions of the API) to predict asthma at school age. In general, a positive API score requires recurrent wheezing episodes during the first three years of life (fewer than three wheezing episodes per year for the loose API and three or more wheezing episodes per year for the stringent API) and one of two major criteria (doctor's diagnosis of eczema or parental asthma) or two of three minor criteria (doctor's diagnosis of allergic rhinitis, wheezing without colds or eosinophil count ≥ 4%; **Table 2.3**)

Caudri *et al.* (88) developed a clinical risk score for preschool children with wheeze or cough to predict asthma at age seven to eight years. They tested a total of 26 candidate predictors assessed between the age of zero and four years, including family history of allergic diseases (six variables), perinatal information (eight variables), environmental factors (four variables), and the child's pattern of symptoms (eight variables). These variables were selected based on published literature as well as availability and were considered to be easy to collect in clinical practice. Only eight variables were significantly associated with asthma between the age of seven and eight years and therefore included in the final predictive model (**Table 2.3** and **Table A.1**); the top five predictors were frequency of wheezing, eczema, history of parental asthma, wheezing without colds and post-term delivery.

Similarly, Pescatore's risk score (108) tested 24 candidate predictors identified through the literature and considered to be readily available in primary care. Candidate predictors were collected in the first three years of life and included demographic (three variables) and perinatal information (two variables), eczema, upper and lower respiratory tract symptoms, triggers and severity of wheeze (17 variables) and family history of allergic diseases (two variables). Only 10 predictors were associated with asthma between the age of six and eight years and therefore included in the final predictive model (**Table 2.3** and **Table A.1**); the most important predictors were shortness of breath, frequent wheeze, wheeze without colds, activity disturbance by wheeze and wheeze/cough triggered by exercise.

In total, 16 different candidate predictors were used to develop the models. All models included a history of self-reported or doctor-diagnosed eczema, the presence of wheeze not associated with colds and a measure of the frequency of early wheezing. Other variables included sex (88, 108), age (87, 108), allergic rhinitis and self-reported or doctor-diagnosed parental asthma or allergies (87, 108). The full list of predictors included in the development models is available in **Table 2.3**. Although a wide variety of predictors was tested, no studies examined the same combination. Aside from blood eosinophilia (87), other potential objective measurements (e.g. total and specific immunoglobulin E [IgE] levels, etc.) were not considered. Diagnosis of asthma and asthma treatment in the past 12 months, were not examined as potential candidate predictors, as they are deemed to be too closely linked to the presence of asthma later in life and would lead to overestimates of the predictive performance of the models. Environmental exposures, such as crowding (defined as density of people in the household), indoor pollution and pets in the house, which are potentially associated with asthma development (141), were not included in the models assessed.

Table 2.3: Candidate predictors explored among development studies included in the systematic review

Candidate predictor	Castro-Rodriguez 2000 (API)	Caudri 2009 (PIAMA)	Pescatore 2014
Sex		✓	✓
Age			✓
Post-term delivery		✓	
Parental education		✓	
Parental inhalation medication		✓	
Wheezing frequency	✓*	✓	✓
Wheezing apart from colds	✓	✓	✓
Infections		✓	
Eczema	✓	✓	✓
Parental asthma	✓		✓
Allergic rhinitis	✓		
Eosinophilia (≥4%)	✓		
Activity disturbance			✓
Shortness of breath			✓
Exercise-related wheeze/cough			✓
Aeroallergen-related wheeze/cough			✓

*Wheezing frequency for stringent API only.

The estimated odds ratio of the predictors included in the final model were reported for both PIAMA (88) and Pescatore's risk scores (108), but confidence intervals were only reported for PIAMA. For the API score, Castro-Rodriguez *et al.* (87) did not report odd ratios (**Table A.1**). The predictors found to be significantly associated with asthma later in life and included in the three predictive models are reported in **Table A.1**. The methodological heterogeneity of the development studies, such as differences in the set of predictors tested and variability in their definitions, did not allow me to compare estimates and perform a meta-analysis.

2.2.3 Outcome definition

Among the development studies, there was considerable heterogeneity in the way researchers defined the outcome (**Table 2.4**)**.** The most commonly used criterion was wheezing in the previous 12 months. To define asthma for the API, Castro-Rodriguez and co-authors used a combination of doctor-diagnosed asthma and at least one wheezing episode in the past 12 months, or three or more

episodes of wheeze regardless of a doctor's diagnosis (87). In the PIAMA study, Caudri *et al.* defined childhood asthma at age of outcome assessment based on any of the following: at least one wheezing episode in the past 12 months, doctor diagnosis of asthma or asthma medication between the ages of seven and eight years (88). The risk score developed by Pescatore *et al.* defined asthma as a combination of wheezing and asthma medication in the past 12 months (108). None of the development studies used objective measurements (e.g. lung function tests) to assess asthma.

Although some validation studies (92, 140) defined outcomes using the same criteria as in the development study, others modified them. To validate the API, one study increased the required number of wheezing episodes in the preceding 12 months from three to four (139), while another study (90) used a different combination of criteria (**Table 2.4**). Hafkamp-de Groen *et al.* did not have data on asthma medication at the age of outcome assessment and validated the PIAMA index using the definition of asthma from Leonardi *et al.* (91, 139).

The age at outcome assessment varied across the development studies, with the API providing asthma prediction from six and up to 13 years (87), while other models examined the outcome only up to age eight years (88, 108). The validation studies assessed asthma at an age that was one or two years different from the age in the development studies.

Table 2.4: Outcome definition across studies

Source of study	Target population	ID	Prediction rule	Age at outcome assessment (years)	Asthma symptoms in the past 12 months[c]	Current wheeze[b]	DDA	Current asthma medications	Other
Development	General	*Castro-Rodriguez 2000*	API	6, 8, 11, 13	-	✔	✔	-	≥3 episod of wheezi in the pas months
	High risk	*Caudri 2009*	PIAMA	7–8	-	✔	✔	✔	
		Pescatore 2014	Risk Score	6–8	-	✔	-	✔	
Validation	General	*Leonardi 2011*	API	7 and 10	-	✔	✔	-	≥4 episod of wheezi in the pas months
	High risk	*Devulapalli 2008*	API	10	✔	✔	✔	✔	Positive exercise t
		Rodriguez-Martinez 2011	API/ PIAMA	5–6	-	✔	✔	✔	≥3 episod of wheezi in the pas months
		Hafkamp-de Groen 2013	PIAMA	6	-	✔	✔	-	≥4 episod of wheezi in the pas months
		Grabenhenrich 2014	Risk Score	8	-	✔	-	✔	

Abbreviations: DDA: doctor's diagnosis of asthma. [a]✔ = part of the asthma definition; dash = not included in the asthma definition
[b]Current wheeze is defined as wheezing or whistling in the chest in the past 12 months. [c]Asthma symptoms in the past 12 months is defined as wh apart from colds or cough or chest tightness. **Castro-*Rodriguez 2000:*** current wheeze (at least one episode of wheeze) and DDA or ≥3 episodes of wh in the past 12 months regardless of DDA***; Caudri 2009:*** current wheeze (at least one episode of wheeze) or DDA or current asthma medications; ***Pes 2014:*** current wheeze and current asthma medications; ***Leonardi 2011:*** current wheeze and DDA or ≥4 episodes of wheezing in the past 12 months reg of DDA***; Devulapalli 2008:*** at least two of the following three criteria: dyspnoea, chest tightness and/or wheezing from 0 to 3 years and/or at 4-10 y age or DDA or use of asthma medication at 0–3 years and/or 4–10 years of age and one of the following criteria: dyspnoea, chest tightness and/or wh in the past 12 months or use of asthma medication in the past 12 months or positive exercise test; ***Rodriguez-Martinez 2011:*** current wheeze and DD episodes of wheezing in the past 12 months regardless of DDA (for API) / current wheeze or DDA or current asthma.

2.2.4 Statistical modelling and predictive performance

Statistical approaches used to develop the predictive models in the original studies included logistic regression analysis and the chi-square test to compare proportions, as well as more advanced penalised regression methods. Details of the statistical methods are reported in **Table A.2**.

In summary, the API classified children as being at low or high risk of having asthma in later childhood based on clinical knowledge and then compared proportions of those two groups of children using a simple statistical test, such as the chi-square test. Caudri *et al.* (88) used a two-step approach, with step one being predictor selection based on univariate analysis followed by multivariable logistic regression, which included statistically significant predictors from the first step. Pescatore *et al.* (108) used a penalized regression method; thus, pre-selection of predictors was not necessary because this method performs variable selection and estimation of regression coefficients simultaneously.

Two of the development studies internally validated their predictive models; Pescatore *et al.* used a leave-one-out cross-validation approach (88, 108), while Caudri *et al*. used a more advanced method (bootstrapping). The maximum proportion of missing data for any potential predictor variables was reported in the PIAMA (<2%) and Pescatore's risk score (≤5.8%), whereas the API reported 19.6% of missing information for the major/minor criteria used. Of the three development studies, only Caudri *et al.* (88) provided information on the method used to handle missing data, and they presented the results from multiple imputation using Multivariate Imputation by Chained Equations (MICE; see **Table A.2**).

In the development and validation studies, the reporting of performance measures varied (**Table A.3**). The discriminative ability measured by the area under the ROC curve (AUC) was reported in five studies [two development (88, 108) and three validation studies (91, 139, 140)] and ranged from 0.63 to 0.83. In general, predictive models with high sensitivity had lower specificity, and discrimination measures were better in studies in high-risk groups with higher prevalence of outcome, compared to those in the general population. Likelihood ratios are an additional measure of predictive performance and indicate the probability that a subject will or will not develop the condition, given the results of a predictive model. Overall, likelihood ratios were not reported in the development studies; only two validation studies calculated them but without providing their

confidence intervals (92, 140). While I could calculate the likelihood ratios for all studies, I could not calculate their confidence intervals (no data reported on standard error), which limits the possibility of comparing them.

The models with higher positive likelihood ratios (a better ability to predict asthmatic children) also had higher negative likelihood ratios. Positive likelihood ratios varied between 1.07 and 7.43, and negative likelihood ratios varied between 0.26 and 0.88. The highest likelihood ratios were reported in the stringent API development study (87).

The performance of the models varied by age at outcome assessment (**Figure 2.2**, **Figure 2.3**, **Figure 2.4** and **Table A.3**). The stringent and loose versions of the API used in the validation studies showed higher sensitivity than in the original study. The highest values were reported in studies carried out in a high-risk population, but similar or lower specificity was found in the external validation studies at all age groups. All externally validated studies of the API reported much wider confidence intervals for the performance measurements than in the development study (**Figure 2.2**, **Figure 2.3** and **Table A.3**).

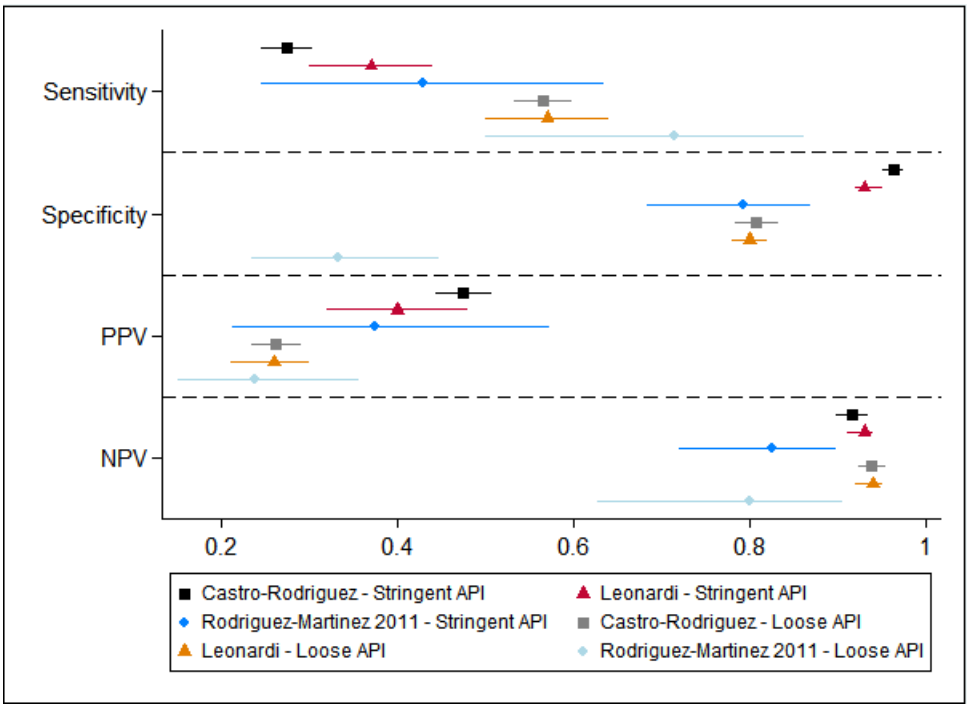


Figure 2.2: Performance measures of API applied to children age five to eight years

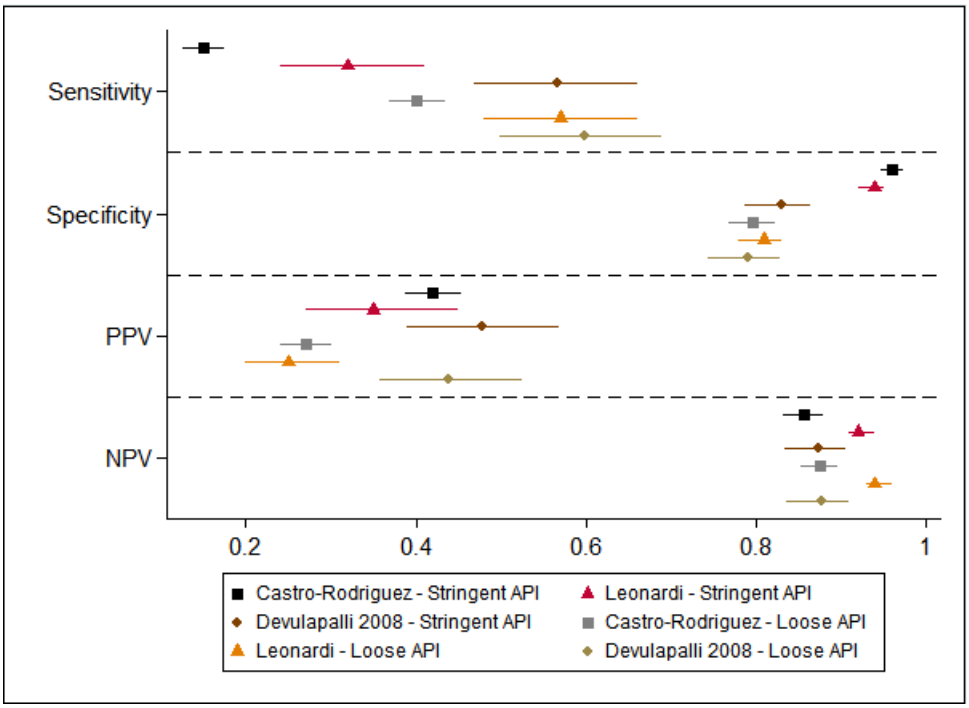


Figure 2.3: Performance measures of API applied to children age 10-11 years

The PIAMA index has been externally validated in two studies, with only one providing most performance measures. The original PIAMA study (88) showed slightly higher sensitivity and positive likelihood ratio, similar specificity and much lower positive predictive value at the age of seven to eight years compared with the study reported by Rodriguez-Martinez *et al.*, which assessed children at age five to six years (92) (**Figure 2.4** and **Table A.3**).

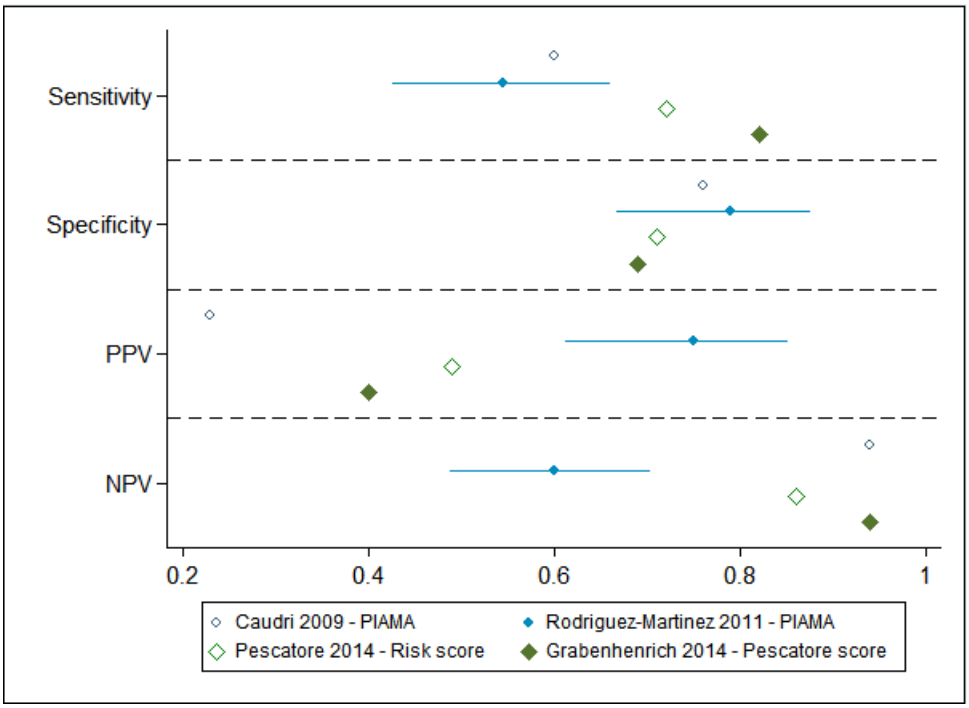


Figure 2.4: Performance measures of the PIAMA and Pescatore's risk score

A predictive score developed by Pescatore *et al.* (108) was validated in a single study (140), which showed, for children age eight years, higher sensitivity, positive likelihood ratio and AUC with a similar specificity and lower positive predictive value than the development study, which assessed childhood asthma between the ages of six and eight years (**Figure 2.4** and **Table A.3**).

2.2.5 Risk of bias assessment

The risk of bias (**Table A.4**) was assessed using the CHARMS checklist (137), with previously reported criteria (86). The risk of participant selection bias was low in all studies but one (92). The risk of bias for predictors and outcome assessment was low in all development and validation studies. Loss to follow up resulted in a moderate attrition in seven of eight studies; the risk of attrition bias was high in the study by Rodriguez-Martinez and co-authors (92), as loss to follow up was higher than 20%, and differences on key characteristics between participants and those who were lost to follow up were not fully described. Four studies provided relevant aspects of the analysis resulting in a low risk of analysis bias (88, 108, 139, 140), and the other four had a moderate risk of such bias (87, 90-92).

2.3 Discussion

The predictive models included in this systematic review showed limited accuracy in forecasting persistent wheeze and asthma at school age. Thus, they have limited clinical usefulness in practice, but other prediction models could be validated in the future. Among all available asthma-predictive models, only the API (87), PIAMA (88) and the Leicester model (108) have been externally validated.

Overall, I showed a discrepancy between the original and validated models in predictive performance, especially when the model was developed in a general population but validated in a high-risk group. The validation studies tended to report higher sensitivities, lower specificities and lower predictive values than the original studies. These results suggest that when predictive models are applied in external validation studies, they have a greater ability to identify children who will develop asthma compared to predictive models established in development studies; however, they also have a propensity to produce prediction errors in forecasting the condition in children whose symptoms will not endure. Because predictive values depend on the prevalence of the condition, models developed in high-risk populations have higher positive predictive value than those developed in general population studies.

Leonardi *et al.* (139), Devulapalli *et al.* (90) and Rodriguez-Martinez *et al.* (92) focused on external validation of the API but in different geographic regions, target populations and clinical settings. The performance of the stringent API in all age groups varies significantly between validation studies, which makes meaningful comparison difficult. This probably reflects the initial population selection,

as the stringent version includes only children with early frequent wheeze. In contrast, the loose version of the API seems to be more reproducible, showing better ability to predict asthma but performing slightly worse in ruling out disease in the external validation studies in all age groups. The positive predictive value of both versions of the API varies among studies, highlighting the importance of defining the population based on early wheezing symptoms, as well as the dependency between this performance measure and disease prevalence.

The variation in the predictive performance between development and validation studies can be explained by artefactual and clinical mechanisms (142). Artefactual mechanisms refer to performance variability arising from flaws in the design and execution of the validation studies; these flaws include misspecification of target population in terms of inclusion criteria for participant enrolment, poor data reporting, application of inappropriate statistical analysis and the lack of a standardised definition of asthma (143). The use of questionnaires to establish asthma diagnosis is an important drawback of the included predictive models. It may lead to overdiagnosis of asthma, as a single episode of wheeze per annum reported by parents without any supportive clinical information might not be appropriate to diagnose asthma accurately. Recent data show that asthma is often overdiagnosed, even in clinical settings (144), with parent-reported histories of cough and wheeze being the main reasons behind inaccurate diagnosis. This suggests that questionnaire data have limited reliability for asthma definition, highlighting the importance of objective measurements and potentially explaining the difference in asthma prevalence in the studies assessed. Objective measurements, such as reversible airflow limitation, raised FeNO and IgE levels, were not used to characterize asthma at school age in either the development or the validation studies.

Possible clinical mechanisms include case-mix variation (145-147), which is a special instance of the spectrum effect. The spectrum effect refers to the variation in sensitivity, specificity and accuracy among different populations and subgroups (148-150), whereas case-mix variation refers to differences, between development and validation studies, in clinical settings, subject characteristics, subject age at outcome assessment, outcome prevalence and/or incidence (145). In particular, the greater the difference between the characteristics of the development and validation populations, the higher the discrepancy in predictive performance, and the poorer the generalisability of findings.

Sensitivity, specificity and likelihood ratios are generally used as benchmarks for comparing model performance, as they are thought to be independent of the prevalence of the outcome (151-153), whereas predictive values vary with the prevalence of the disease. However, several studies have shown that sensitivity and specificity, together with likelihood ratios, might also vary with disease prevalence (143, 154, 155). Disease prevalence is often predetermined by inclusion criteria, target population specifications and outcome definitions; for example, when individuals are selected based on their symptoms and risk profile (a 'high-risk' population), the prevalence is higher than in the general population (e.g. the Tucson Children's Respiratory Study is a general population and, as expected, the prevalence of asthma is lower than that in the high-risk validation studies from Norway and Colombia). Although it is important to assess the predictive performance of a model in a new sample of individuals, from different locations and clinical settings, the validation cohort should be homogeneous with that of the development study in order to reduce variation in performance, to facilitate comparison among between models and to establish the generalisability of the results.

Differences in age at outcome assessment can also lead to case-mix variation and change in predictive performance. Prediction is usually challenging, and predictive performance might be limited when the outcome is assessed in the future using information collected earlier in life. Most studies of childhood asthma prediction aimed to develop and validate models to predict the occurrence of asthma between six and 10 years of age by using predictors collected at age three to five years, making prediction potentially relatively simple and reliable. The validation studies that assessed asthma at ages similar to those of their development counterparts had comparable results, but no studies aimed to develop and externally validate asthma predictive models after age 13 years.

At present, asthma is considered an umbrella term (156) unifying different conditions that share common clinical features, such as cough and wheeze, shortness of breath and bronchial obstruction (157). Although distinct wheezing and asthma endotypes have been proposed, existing guidelines based on clinical symptoms and predictive models still assume a single disease (16). This discrepancy might partially explain the inaccuracy of asthma prediction. Future work should consider endotypes to a larger extent in the development of predictive models.

Another important issue that potentially affects predictive performance is individual genetic predisposition to disease development. This important factor was not taken into account in the studies included in this systematic review. With more genome-wide association studies (GWAS) being conducted, genetic variants that predispose individuals to asthma (158) have been discovered; for example, a recent meta-analysis of GWAS reported seven genetic risk loci (HLA-DQ, IL33, ORMDL3/GSDMB, IL1RL1/IL18R1, IL2RB, SMAD3 and TSLP) for asthma (13, 74). Considering these genetic factors in the development of future predictive models may lead to improved predictive performance. However, this information is currently rarely available at a clinical visit, and its inclusion would limit the routine use of a predictive model in clinical practice.

To better assess the accuracy of existing asthma-prediction models, future research should collect information by using well-standardised questionnaires and clinical tests, to allow for better harmonisation of predictor and outcome definitions. Collaborative international, multicentre efforts would provide opportunities to increase sample size and to develop predictive models using data from subsets of studies and then validate them in the remaining ones. More studies performed in general and high-risk populations in parallel will allow for meaningful evaluation of model performance. Clinical tests, including lung function tests, skin-prick tests, allergen-specific IgE, fractional exhaled nitric oxide (FeNO) and bronchodilator response, could also be used to increase the capability of a predictive models to correctly identify children at high risk of asthma in adolescence. Moreover, external validation study populations should be carefully selected to have characteristics similar to those of the study populations from which the predictive models were developed, in order to have low heterogeneity within subpopulations.

2.4 Conclusion

This systematic review evaluates the performance of externally validated predictive models for childhood asthma. This topic is of high importance to practicing physicians, as many parents want to know the probability that their child will develop asthma, as well as to researchers, because screening is crucial to establish future experimental and epidemiological studies.

The number of validated instruments is limited, and they provide poor predictive accuracy, with most performance variation in sensitivity and positive predictive value. Variation in predictive

performance between development and validated asthma models can be explained by artefactual and clinical mechanisms, such as differences in recruitment criteria, definitions of the target population, age at outcome assessment, prevalence of the condition, as well as differences in data collection methods due to different surveys administrated and/or a misinterpretation of respiratory condition by parents and/or doctor and overuse of inhaled steroid medications.

Better predictive ability can be achieved by studying general and high-risk groups in parallel in the same study, to assess an instrument's predictive performance in full. The asthma predictive models that have not been externally validated should be evaluated to ensure their reliability. Larger cohorts are needed so that more predictors can be included in the models so that more advanced statistical methods can be used. More work on standardisation of predictors and outcome assessment needs to be done. Until such work is completed, currently available asthma prediction models cannot be recommended for use in clinical practice and research studies.

Chapter 3 – Asthma in Ashford: a 20-year follow-up

The Ashford study is a prospective birth cohort study that recruited pregnant women who gave birth to 642 infants in Ashford, Kent between 1992 and 1993 (159) (**Figure 3.1**). Specifically, all newly pregnant women attending an antenatal visit to one of three general practitioner surgeries in Ashford were approached to participate. Overall, 710 women were invited, and 658 (93%) enrolled in the study. In the intervening period, a series of follow-up surveys identified children who had developed allergic and other wheezing disorders. Questionnaires were administered to the mothers during pregnancy, and then annually around their child's birthday. Environmental measurements were collected from the babies' homes soon after birth. Additional environmental data were collected at each annual visit, along with health information, such as wheezing and other respiratory symptoms, other allergies and family history (including skin prick tests for the parents). A total of 499 children (approximately 80% of the original cohort) provided data at the last follow-up, at age 14 years.



Figure 3.1: Children of the Ashford cohort

In 2013, the Ashford cohort joined the Study Team for Early Life Asthma Research (STELAR) consortium (116) together with four other UK asthma birth cohorts (ALSPAC, Isle of Wight, MAAS and the Study of Eczema and Asthma To Observe the influence of Nutrition [SEATON]) (95, 160-164). In March 2015, a new follow-up study using a short, self-completed, paper or web questionnaire was planned to collect information from the Ashford cohort children, then aged 20 years. The new survey collected data from childhood to adolescence that would be comparable across the five STELAR asthma birth cohorts. Similar follow-ups were planned for the other birth cohorts, and the information collected was incorporated and harmonised with that collected in earlier surveys of each study population; details on the harmonized information can be found in Chapter 4 (Data sources, paragraph 4.3). Information from the subjects in the Ashford cohort (at age of 20 years) as well as data collected among the STELAR cohorts at a similar age is used to define asthma in adolescence (between 15 and 20 years of age) as an outcome for predictive models; details on this are reported in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years) and Chapter 6 (Further investigations using the Bayesian model).

The design of the new survey was based on the International Study of Asthma and Allergies in Childhood (ISAAC) questionnaire (165-168), which was established to standardize asthma and allergic disease information collated in epidemiological studies, as well as questionnaires administered in the other UK birth cohorts during the intervening period. This survey design and development process allowed me to create a survey consistent with the other UK asthma cohorts and to assess the content validity as well as internal consistency of the Ashford survey.

The response rates of epidemiologic studies have declined in the past 30 years (169, 170), a trend that is more evident in recent years (171-173). A low response rate reduces the sample size and can result in inconsistency between respondents and non-respondents in terms of characteristics measured. Consequently, there is a risk of bias in estimates due to differences between people who responded to the survey and those who did not, leading to non-representative results of the study population. As non-response bias can affect the interpretation of results, the use of methods that maximise the response rate is very important (174). I used both paper- and web-based questionnaire formats in order to increase the response rate and to limit non-response bias. The paper

questionnaire has been the most popular epidemiological method to collect information from research studies so far, but the increasing use of the Internet has made web-based surveys more accessible, particularly to young people. Online methods have lower cost (175-178), which can be important in large population studies, but response rates have been reported to be lower for web-based questionnaires than for paper-based methods (179-181).

The main differences between web-based and paper questionnaires are their cost, their presentation and the speediness of obtaining responses. Web-based questionnaires have numerous advantages, including the creation of a tailored question path through the questionnaire that varies based on a respondent's previous answers, inexpensive emails and reminders to participants, standardized and validated responses, and data can be gathered quickly from a large number of respondents and downloaded to a spreadsheet for easy processing and analysis (182-184). None of those is possible for paper questionnaires. However, web questionnaires have issues concerning their difficulty to use by people who lack computer expertise and differences between monitors and browsers such that questions might be visualized and presented differently from one participant to another (182-184). These problems could impact the response rate and introduce bias.

This chapter describes the aims of the Ashford birth cohort study, its structure and the characteristics of the study population, with a brief description of exposure measurements collected over the years since its inception, in particular the methods used to collect information when the children were 20 years of age. Statistical analyses compared the differences between respondents and non-respondents as well as children with and without asthma at age 20 years in terms of observable characteristics. Finally, I calculated the proportion of subjects who completed the paper- versus web-based questionnaires and compared their characteristics.

3.1 Methods

3.1.1 The Ashford study

An online questionnaire was developed using Qualtrics (https://imperial.eu.qualtrics.com), a protected online survey system offered by Imperial College. The survey (**Figure B.1**) focused on questions related to asthma and its impact on daily life; it included four main sections:

- Personal information providing an overview of general characteristics of the participants
- Home environment, including questions about the presence of mould, condensation and pets in the home
- Health questions on wheezing, asthma, skin problems, treatments and triggers
- Smoking and leisure activities

Before conducting the survey, the questionnaire was pilot tested to identify and amend problematic items related to its content, wording, format, length and instructions. Designing and piloting questionnaires is a complex and time-consuming process, but the quality of the information collected is determined by the quality of the questionnaire (185, 186). Therefore, a draft of the questionnaire was first sent to colleagues for comments, and the survey was then piloted on a small sample of people (n=10) who have characteristics similar to those of the target population in terms of age, sex, ethnicity and education. Decisions were made based on their comments, including enlarging the number of specific questions; improving the language and wording to make the questions less abstract, more precise and simpler; changing the logical position of some items and adding question-skip logic, which directs respondents through customised paths. Since this research proposal involved human participants, I submitted to the Head of Department and the Imperial College Research Ethics Committee (ICREC) an application form along with all supporting documentation, such as the protocol, sample letter to possible participants (**Figure B.2**), the paper questionnaire (**Figure B.1**), participant information sheet and consent form. They reviewed and approved the study, as no ethical issues were found (http://www.imperial.ac.uk/research-ethics-committee/).

Families' details, such as home address and postal code, were collated from 489 children who had participated in the previous follow-up; only about 350 children's parents had provided an e-mail address. Therefore, once ethical approval had been obtained, the paper questionnaire was sent to the 489 children, and the web-based questionnaire was sent to the 350 parents who gave their consent to be contacted via email; the parents were asked to invite their (cohort) children to complete the online questionnaire. Moreover, personalised web links were added to the paper questionnaires, so that all participants had the option to complete the questionnaire using either

format. As an incentive for completing and returning the survey, five £50 prizes were offered on the basis of a lottery among respondents.

The questionnaires were first dispatched in late February 2015; the response rate was about 25%. In order to engage more participants, up to two reminders were sent. First and second reminders were sent in May and July 2015, respectively. Of the 489 eligible participants, 72 (15%) proved impossible to trace. Of the 417 remaining participants, 245 (60%) returned completed questionnaires. The average cost of the paper questionnaire and its delivery was about £10 per subject (including cost of envelopes, stamps, papers, cartridge for printing questionnaires and invitation letters and returned envelopes) compared to the minimal cost of the online questionnaire.

Data collected were entered and checked for input errors using the EpiData Manager software. EpiData is Windows-based public domain software that provides instruments for database design, validation and data entry; this software allows information to be entered in a simple text form and then converts it into a database. I created an EpiData project to define the data structure and added metadata (labels and definitions) to control variable levels by using restrictions on values, range and filtered questions. A random sample of the data (20%) was double-entered in order to identify typing errors. Data were subsequently exported for quantitative analysis.

3.1.2 Statistical analysis

This section describes a simple statistical analysis of the Ashford follow-up data at age 20 years. Reported analyses are based on data from 245 respondents and 172 non-respondents, defined as subjects whom I believe received the questionnaire but did not return it.

The primary aim of these analyses is to evaluate non-response bias that occurs when non-respondents differ from respondents in observable characteristics for which information is available (sample selection bias). In order to identify features that were independently associated with participation in the follow-up survey, a multivariable logistic regression analysis was performed to estimate the effect of each explanatory variable on response, while controlling for the other variables in the model.

The secondary aim is to compare differences between subjects with and without asthma at age 20 years, including their socio-demographic characteristics, indoor and outdoor environmental exposures, respiratory symptoms, general health conditions and leisure activities. A multivariable logistic regression was performed to identify characteristics that were associated with asthma at age 20 years, while controlling for the other variables.

The third aim is to assess the response rates for paper- and web-based questionnaires and to compare sociodemographic characteristics and general health conditions of participants who used a paper questionnaire versus those who used a web-based questionnaire.

The administered questionnaires (web-based and paper), as well as detailed definitions of variables based on questions asked to study participants, are reported in the appendix (**Figure B.1**). Continuous outcomes were summarised by using their median values (with interquartile ranges), as the distribution of such variables was non-normal. Comparisons of continuous variables in the two groups were performed using the Mann-Whitney U test for non-normal variables. Categorical variables were summarised using frequencies and percentages. Differences in categorical variables between outcomes were analysed using the chi-square test or Fisher's exact test, whichever was appropriate. Odd ratios (ORs) together with 95% confidence intervals and p-values for comparison between groups for each outcome were reported, and statistical significance was evaluated using a two-sided significance threshold of 0.05.

3.2 Results

3.2.1 Differences between respondents and non-respondents

Respondents and non-respondents were similar with regard to most sociodemographic factors and asthma-related symptoms at age 14 years, but they differed significantly in terms of sex and doctor's diagnosis of asthma at any time. The proportions of females and subjects who had not had a doctor's diagnosis of asthma or wheezing in the past 12 months were greater among respondents compared to non-respondents (**Table 3.1**). Specifically, the odds ratio of being a respondent for this follow-up study was 0.55 (95% confidence interval 0.37 to 0.81) for males compared to females, and 0.58 (95%

confidence interval 0.36 to 0.94) for participants who had a doctor's diagnosis of asthma compared to those who never had this diagnosis.

Table 3.1: Characteristics of subjects at age 14 years by participation status in the 20-year follow-up

Variable	Non-respondent N = 172 n (%)	Respondent N = 245 n (%)	OR[a] (95%CI[b])	*P*-value[c]
Sex (Male)	103 (59.9)	110 (44.9)	0.55 (0.37,0.81)	**0.004**
Ethnicity (Caucasian VS Others)	161 (93.6)	235 (95.9)	0.97 (0.16,5.89)	>0.99
Wheezing in the past 12 months	28 (16.3)	32 (13.1)	0.77 (0.45,1.34)	0.435
Itchy rash in the past 12 months	47 (27.3)	67 (27.4)	1.00 (0.65,1.55)	>0.99
Ever diagnosed with asthma	44 (25.6)	41 (16.7)	0.58 (0.36,0.94)	**0.037**
Ever diagnosed with hayfever	57 (33.1)	92 (37.6)	1.21 (0.81,1.83)	0.411
Ever diagnosed with eczema	58 (33.7)	99 (40.4)	1.33 (0.89,2.00)	0.205

[a]Odds ratio (OR) of participation (response) for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values (p<0.05).

Similar to this unadjusted analysis, the multiple regression model showed an association between responsiveness and sex as well as doctor's diagnosis of asthma at any time (**Table 3.2**).

Table 3.2: Results from the multiple logistic regression analysis on participation at age 20 years

Variable	OR[a] (95%CI[b])	*P*-value[c]
Sex (Male)	0.59 (0.39,0.89)	**0.012**
Ethnicity (Caucasian VS Others)	0.90 (0.14,5.78)	0.913
Wheezing in the last 12 months	1.04 (0.52,2.09)	0.903
Itchy rush in the last 12 months	0.80 (0.47,1.35)	0.400
Ever diagnosed with asthma	0.53 (0.29,0.98)	**0.042**
Ever diagnosed with hayfever	1.46 (0.92,2.33)	0.111
Ever diagnosed with eczema	1.56 (0.96,2.54)	0.074

[a]Odds ratio (OR) of participation (response) for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values (p<0.05).

3.2.2 Definition of active asthma at age 20 years

Subjects at age 20 years were considered to have active asthma if they had two or more positive answers to the following items: a) at least one episode of wheezing in the past 12 months; b) current asthma treatment prescribed by a doctor; and c) a doctor's diagnosis of asthma at any time.

Overall, 34 (13.9%) respondents who reported symptoms, use of asthma medication or both at age 20 years were thus defined as having active asthma (**Table 3.3**). Most subjects had a previous doctor's diagnosis of asthma (20.8%), 19.6% of subjects reported current wheezing symptoms, and only 24 subjects (10%) reported use of medication.

Table 3.3: Prevalence of features of asthma at the age of 20 years

	Wheezing at least once		**Asthma treatment**		**Doctor's diagnosis of asthma**		**Asthma (positive on ≥2 items)**	
Age (years)	No.	%	No.	%	No.	%	No.	%
20	48	19.6	24	10.0	51	20.8	34	13.9

3.2.3 Differences between asthmatics and non-asthmatics at the age of 20 years

Respondents with and without asthma did not differ significantly in terms of sociodemographic characteristics. Respondents with asthma were more likely to be smokers and have their own children, but the differences were not statistically significant (**Table 3.4**). A weak association was found for employment status; the odds of having asthma at 20 years is three times higher in respondents who have a part-time job compared to those who work full time, although the confidence interval was wide, ranging from 1.16 to 8.05.

Table 3.4: Characteristics of the study population by presence of asthma at age 20 years

Variable	Non-asthma N = 211 n (%) or median (IQR)	Asthma N = 34 n (%) or median (IQR)	OR[a] (95%CI[b])	*P*-value[c]
Sex (Male)	94 (44.6)	16 (47.1)	1.11 (0.54,2.29)	0.919
Ethnicity (Caucasian VS Others)	202 (95.7)	33 (97.1)	0.33 (0.03,3.71)	0.372
Education				0.762
Full-time education	97 (46.0)	15 (44.1)	Reference	
Part-time education	11 (5.2)	2 (5.9)	1.62 (0.31,8.35)	
Not in any education	103 (48.8)	17 (50.0)	1.07 (0.51,2.25)	
Paid job				0.065
Full-time paid work	84 (40.0)	7 (20.6)	Reference	
Part-time paid work	56 (26.3)	14 (41.2)	3.05 (1.16,8.05)	
Not working in a paid job	71 (33.7)	12 (35.3)	2.03 (0.76,5.43)	
Own children	6 (2.8)	2 (5.9)	0.48 (0.09,2.47)	0.314
Smoking in the past 30 days	42 (20.0)	9 (26.5)	1.50 (0.65,3.47)	0.469

[a]Odds ratio (OR) of asthma for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test and Mann-Whitney U test. There were no statistically significant *p*-values ($p<0.05$).

Table 3.5 shows that a significantly higher proportion of respondents with asthma at age 20 years had episodes of dry cough at night and medical diagnosis of hayfever or eczema, compared to those without asthma. The odds of having asthma at age 20 years was 4.95 (95% confidence interval 1.96–12.48) times greater in respondents reporting hayfever, 8.13 (3.67–18.02) times higher in respondents reporting dry cough at night without a cold and 2.42 (1.15–5.11) times higher in respondents who had a doctor's diagnosis of eczema compared to those who never had these conditions.

Table 3.5: Respiratory and other symptoms by presence of asthma at age 20 years

Variable	Non-asthma N = 211 n (%)	Asthma N = 34 n (%)	OR[a] (95%CI[b])	*P*-value[c]
Itchy rash in the past 12 months	63 (29.9)	15 (44.1)	1.93 (0.92,4.07)	0.122
Ever diagnosed with hayfever	100 (47.4)	27 (79.4)	4.95 (1.96,12.48)	**<0.001**
Dry cough at night without a cold	27 (12.8)	18 (52.9)	8.13 (3.67,18.02)	**<0.001**
Ever diagnosed with eczema	75 (35.5)	19 (55.9)	2.42 (1.15,5.11)	**0.029**
Food allergy	24 (11.4)	5 (14.7)	1.38 (0.49,3.90)	0.564

[a]Odds ratio (OR) of asthma for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values ($p<0.05$).

Table 3.6 shows no statistically insignificant associations between environmental factors and asthma at age 20 years. A significantly greater proportion of respondents with asthma reported physical health and limitation of social activities (**Table 3.7**). The odds of having asthma at age 20 years was 3.37 (95% confidence interval 1.43–7.93) times greater in respondents reporting physical health and limitation of social activities compared to those who never had these conditions.

Table 3.6: Indoor and outdoor environmental risk factors by presence of asthma at age 20 years

Variable	Non-asthma N = 211 n (%) or median (IQR)	Asthma N = 34 n (%) or median (IQR)	OR[a] (95%CI[b])	*P*-value[c]
Problem with damp or condensation	59 (28.0)	7 (20.6)	0.73 (0.30,1.78)	0.626
Problem with mould	54 (25.6)	7 (20.6)	0.87 (0.35,2.14)	0.931
Pets at home	143 (67.8)	19 (55.9)	0.59 (0.28,1.24)	0.229
Traffic on the road live (Not heavy traffic VS Heavy traffic)	170 (80.6)	23 (67.7)	0.55 (0.25,1.26)	0.231
Number of people living in the household	4 (2)	4 (1)	1.02 (0.79,1.31)	0.399
Years in your house	15 (13)	13 (13.8)	0.98 (0.94,1.03)	0.660

[a]Odds ratio (OR) of asthma for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test and Mann-Whitney U test. There were no statistically significant *p*-values ($p<0.05$).

Table 3.7: Health and physical activity by presence of asthma at age 20 years

Variable	Non-asthma N = 211 n (%)	Asthma N = 34 n (%)	OR[a] (95%CI[b])	*P*-value[c]
General health				**0.005**
Excellent	52 (24.7)	2 (5.9)	Reference	
Very Good	87 (41.2)	10 (29.4)	2.99 (0.63,14.17)	
Good	49 (23.2)	17 (50.0)	9.02 (1.98,41.07)	
Fair or Poor	23 (10.9)	5 (14.7)	3.90 (0.61,25.09)	
Physical health interfered with social activities	24 (11.4)	5 (14.7)	3.37 (1.43,7.93)	**0.005**
Frequency of exercise				0.358
Daily	75 (35.5)	7 (20.6)	Reference	
Weekly	75 (35.5)	14 (41.2)	2.00 (0.76,5.23)	
Monthly	17 (8.1)	4 (11.8)	2.52 (0.66,9.59)	
Less than monthly	17 (8.1)	1 (2.9)	0.63 (0.07,5.47)	
Once or twice a year	11 (5.2)	3 (8.8)	2.92 (0.66,13.01)	
Never	13 (6.2)	3 (8.8)	2.47 (0.57,10.81)	

[a]Odds ratio (OR) of asthma for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values ($p < 0.05$).

In the multiple logistic regression analysis, because the sample size was small compared to the number of variables I wanted to test, only variables with *p*-value ≤ 0.1 in the univariate analysis were considered. Results showed a relationship between asthma and respiratory symptoms, such as hayfever and dry cough at night without cold, as well as a strong association with physical health leading to limitation of social activities (**Table 3.8**). The odds of having asthma at age 20 years in respondents with a dry cough at night was 8.53 (95% confidence interval 3.14–23.20) times higher than those without the condition, while holding the other variables at fixed values. The odds of having asthma at age 20 years in respondents who had a doctor's diagnosis of hayfever or difficulties with social activities was around 4.5 times higher than those who do not experience those conditions. Confidence intervals for these estimates are wide, suggesting high variability, due to the small sample size.

Table 3.8: Results from the multiple logistic regression analysis on asthma at age 20 years

Variable	OR[a] (95%CI[b])	*P*-value[c]
Paid job		
Full-time paid work	Reference	
Part-time paid work	2.55 (0.76,8.54)	0.128
Not working in a paid job	1.21 (0.37,3.96)	0.752
Ever diagnosed with hayfever	4.58 (1.43,14.59)	**0.010**
Dry cough at night without cold	8.53 (3.14,23.20)	**<0.001**
Ever diagnosed with eczema	1.00 (0.38,2.59)	>0.99
General health		
Excellent	Reference	
Very Good	1.50 (0.29,7.77)	0.630
Good	1.79 (0.33,9.90)	0.502
Fair or Poor	0.33 (0.03,3.50)	0.360
Physical health interfered with social activities	4.54 (1.21,17.06)	**0.025**

[a]Odds ratio (OR) of asthma for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values ($p<0.05$).

3.2.4 Differences between paper and web questionnaire

Overall, 89 subjects completed the questionnaire online (36.3%), while the remaining 156 preferred the paper questionnaire (63.7%). These two groups of respondents were similar with regard to socio-demographic characteristics at age 20 years but differed significantly in terms of general health and completion methods. Respondents who used the online questionnaire were more likely to report general health issues, to be in full-time education and not working in a paid job and to have completed the questionnaire on their own, compared to respondents who used the paper questionnaire (**Table 3.9**). The strongest significant association showed that the odds of respondents completing the questionnaire on their own was 3.57 (95% confidence interval 1.43–8.91) times greater in subjects who used the online survey compared to those who filled out the paper survey.

Table 3.9: Characteristics of respondents at age 20 years by type of questionnaire

Variable	Online N = 89 n (%)	Paper N = 156 n (%)	OR[a] (95%CI[b])	*P*-value[c]
Sex (Male)	41 (46.1)	69 (44.2)	0.93 (0.55,1.57)	0.885
Ethnicity (Caucasian VS Others)	82 (92.1)	153 (98.1)	3.73 (0.33,41.77)	0.285
Education				0.082
Full-time education	50 (56.2)	62 (39.7)	Reference	
Part-time education	2 (2.2)	8 (5.1)	3.23 (0.66,15.88)	
Not in any education	37 (41.6)	83 (53.2)	1.81 (1.06,3.10)	
Paid job				0.157
Full-time paid work	28 (31.5)	63 (40.4)	Reference	
Part-time paid work	24 (27.0)	45 (28.8)	0.83 (0.43,1.62)	
Not working in a paid job	37 (41.5)	46 (29.5)	0.55 (0.30,1.03)	
General health				**0.035**
Excellent	18 (20.2)	36 (23.1)	Reference	
Very Good	27 (30.3)	70 (44.9)	1.30 (0.63,2.66)	
Good	34 (38.2)	32 (20.5)	0.47 (0.22,0.99)	
Fair or Poor	10 (11.3)	18 (11.5)	1.14 (0.40,3.28)	
Physical health interfered with social activities	11 (12.4)	23 (14.7)	1.24 (0.58,2.69)	0.578
General health compared to 5 years ago				0.296
Much better	20 (22.5)	27 (17.3)	Reference	
Much worse	25 (28.1)	35 (22.4)	1.04 (0.48,2.25)	
Has not changed	44 (49.4)	92 (59.0)	1.55 (0.78,3.06)	
Illness or disability	12 (13.5)	25 (16.0)	0.79 (0.38,1.67)	0.667
Completion on your own	83 (93.3)	124 (79.5)	3.57 (1.43,8.91)	**0.007**

[a]Odds ratio (OR) of asthma for each explanatory variable.
[b]95% confidence interval.
[c]*P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values ($p < 0.05$).

3.3 Discussion

The Ashford study is a follow-up study of locally representative children for whom information has now been collected throughout a 20-year study period. Cohort studies such as this, in which a group of people are followed prospectively over a period of time, are often employed in epidemiology in order to assess disease causation by adjusting for multiple risk factors that might be associated with the clinical outcome. They enable researchers to collect information on multiple health outcomes simultaneously, consider several risk factors as well as exposures at different time points, assess the

relationship between risk factors and the development of condition(s) and compute rates of disease in exposed and unexposed participants over time (187-189).

However, the Ashford study is not representative of the whole UK child population, as mostly Caucasian children were recruited and followed-up, making the results difficult to reproduce in populations with different sociodemographic characteristics and clinical settings. Sample representativeness is crucial to ensure that the right mix of subjects included into the sample reflects the entire population and to produce findings that are generalisable to other countries and/or municipalities.

In this chapter, I investigated non-response bias, tested the differences between respondents with and without active asthma and compared survey-completion methods (web- and paper-based questionnaires). The statistical analyses, both univariate and multivariate, showed that females and young adults without asthma and wheezing symptoms were more likely to return a completed questionnaire. As expected, the presence of asthma at age 20 years was strongly associated with respiratory symptoms, such as hayfever, eczema and dry cough at night without a cold. The data also suggested possible associations of asthma with smoking, exposure to air pollution or pet allergens, and with employment status, although most of these associations did not reach statistical significance. Subjects with asthma were less likely to work full time compared to those without asthma; this may be due to the general health issues that children with asthma experience. In fact, respondents with asthma reported more difficulties in daily activities, including social activities, as well as reduced frequency of exercise, compared to non-asthmatics.

Different completion methods were used to increase the response rate and to limit non-response bias. The majority of responders (64%) completed a paper questionnaire; reasons for this may include lack of expertise in using electronic devices and differences in monitors or browsers leading to incorrect visualization of the questions. Sociodemographic characteristics, such as sex and ethnicity, were not associated with the type of questionnaire used. However, a higher proportion of respondents who completed the questionnaire online reported more severe health problems, and they were more likely to be in full-time education, be unemployed and complete the questionnaire on their own, when compared with those who used the paper-based questionnaire.

The information on wheezing symptoms and use of asthma medications in the past 12 months collected in this 20-year follow-up study will be used for future analysis in this thesis. This information is required to define asthma in adolescence (that is, age between 15 to 20 years) and to develop appropriate asthma prediction models; details of the prediction analyses are reported in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years) and Chapter 6 (Further investigations using the Bayesian model).

Chapter 4 – Data sources

Recent commentaries have highlighted the importance of the collaborative use of information for next-generation research in epidemiology and public health (190, 191). Similar information is often available from multiple studies and other sources, and if it is to be integrated then it needs first to be combined. The harmonisation process consists of pooling similar information across different studies and sources to create a single, cohesive dataset that is used for statistical analyses. This process establishes the final form of the dataset and enables larger sample sizes that enhance both statistical power and generalisability to different populations (191-194). Although this process augments the quality of research and maximises its impact, managing and harmonising large datasets can be challenging and time consuming (195). Internet-based networking technologies as well as collaboration between researchers, including clinicians, statisticians, computer scientists, engineers and biologists, should ensure effective solutions to these issues and result in a step forward for future research.

The main dataset used herein contained integrated information from the following five UK asthma cohorts in the Study Team for Early Life Asthma Research (STELAR) consortium (116): the Avon Longitudinal Study of Parents and Children (ALSPAC), the Ashford study, the Isle of Wight study (IoW), the Manchester Asthma and Allergy Study (MAAS) and the Study of Eczema and Asthma To Observe the influence of Nutrition (SEATON) (95, 160-164). These cohorts include children followed from birth to age 20 years in five areas of the UK, namely, Bristol (ALSPAC), Ashford (Kent), Isle of Wight (IoW), Manchester (MAAS) and Aberdeen (SEATON) – see section 4.1.1 for details. To facilitate the harmonisation process, data managers for each cohort provided comprehensive descriptions of their datasets and uploaded them to the Asthma e-Lab, an innovative platform developed to store information from the five cohorts (see section 4.1.2) (116). I then catalogued, annotated and uploaded information gathered in the Ashford study.

This process necessitated detailed specifications of the assembled information in order to achieve a high degree of comparability among the five cohorts (see section 4.2). Once the STELAR datasets were available in e-Lab, I identified compatible information from similar follow-up studies of the five cohorts (see section 4.3.1), and then downloaded, transformed (see section 4.3.2) as well as matched (see section 4.3.3) it to create the final harmonized STELAR dataset used for the statistical analyses reported in the following chapters.

4.1 STELAR consortium

In 2013, the Ashford cohort joined the STELAR consortium (116), a collaborative network of all UK birth cohorts studying asthma (Ashford, ALSPAC, IoW, MAAS, and SEATON) and the North West Institute for Bio-Health Informatics and Microsoft Research in Cambridge. The main aim of STELAR, which was supported by a Medical Research Council project grant, is to explore the natural transition of asthma from childhood to adulthood. Although similar collaborative studies have been done in the past, none has gathered so much clinical and environmental information from different longitudinal cohorts, the subjects of which are now at a similar age. All cohorts in the STELAR consortium investigated asthma and other allergic disorders, their development during childhood and their risk factors. Data collected included the following:

- Questionnaires based on ISAAC and local protocols: parental occupation, history of allergic diseases in parents and siblings, smoking status and passive smoking, pets in the house, environmental factors, quality of life, allergies and respiratory symptoms, such as frequency of wheezing, cough and runny nose
- Lung function tests: spirometry, methacholine challenge, bronchodilator and exercise responses
- Clinical assessment: eczema, hayfever, anthropometric measurements, skin prick tests (sensitisation to mite, cat, dog, grass, tree, milk, egg, peanut) and IgE measurements
- Biological samples: blood and saliva samples for DNA extraction

4.1.1 Study populations

4.1.1.1 ALSPAC

The Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort (163, 164) was designed to investigate the relationship between an individual's genotype and environmental exposures on child (and parental) health and development. The study population was drawn from the part of the County of Avon that was within the South West Regional Health Authority; this area has a population of one million and includes the city of Bristol, a mixture of rural areas, suburbs and other moderate-size towns. The study recruited pregnant women with an expected date of delivery between 1st April 1991 and 31st December 1992 through antenatal clinics in the former County of Avon. In total, recruited women gave birth to 14,451 babies who have been followed up from birth to age 18 years. Questionnaires for the ALSPAC study were first designed and piloted on around 100 parents, and then the information collected was validated with family medical records and clinical examinations. Information was collected from early pregnancy using a variety of sources, including self-completion questionnaires to mothers, partners and their children; medical, educational and other records; environmental measurements in sub-samples of homes, including air pollutants magnetic radiation and noise; clinical visits; and biological samples from mothers, partners and study children. Assessments of children at age seven years included anthropometric measurements, such as weight; height; head, chest, waist and hip circumferences; clinical examination of the skin for eczema; serum-specific IgE and skin prick tests to common allergens. At age of eight and a half, study children underwent lung function tests to measure forced expiratory volume in one second (FEV_1) as well as forced vital capacity (FVC) and methacholine bronchial challenge. Biological samples from mothers who gave consent and their children were collected and stored. Additionally, a blood sample and dietary diary were also collected at age seven years so that DNA could be banked and to provide information on genetic variation.

4.1.1.2 Ashford

The Ashford study (160) was designed specifically to assess the relationships between atopy and asthma and the intensity of allergen exposure during infancy. Ashford, Kent is 100 km southeast of London and 30 km from the coast; it is typical of medium-sized towns in the region, with a population

of approximately 80,000 people and employment mainly in light industry. This study recruited pregnant women at antenatal visits; they gave birth in Ashford to 642 children (95% of those eligible) between 1992 and 1993. When study children were age eight weeks, exposure measures were obtained, including concentrations of house dust mite and cat allergens in the dust from the living room. Questionnaires that enquired into wheezing symptoms in the past year, allergic diseases, respiratory, nasal as well as dermatological symptoms and environmental exposures, were administered to children's mothers (in person) at the (child's) ages of one, two, three, four, five and eight years. Study participants completed the questionnaires on their own at the ages of 14 and 20 years. Additional information collected over the study period included history of parental and sibling allergies, cigarette smoking in the home, pet ownership and occupation. Peak expiratory flow (PEF) before and after exercise and salbutamol, and skin prick tests were performed at five, eight and 14 years of age. Additionally, spirometry measures, such as FEV_1 and FVC were performed at age 14 years. Mouth swabs and blood samples were obtained at school age and stored for DNA extraction. Further details of the Ashford study are reported in Chapter 3 (Asthma in Ashford: a 20-year follow-up study).

4.1.1.3 Isle of Wight

The Isle of Wight (IoW) (95) cohort was established to assess the prevalence of allergic sensitization and its clinical manifestations in an unselected population, to explore the natural history of allergic conditions, to define the heterogeneity of asthma and other allergic conditions across the life course, to identify risk factors for asthma and similar diseases and to investigate gene-environment interactions in the development of asthma and allergic conditions. The Isle of Wight has a total area of 147 square miles and is between two and five miles from the coast of Hampshire; it has a population of around 141,000. The Isle of Wight Health Authority and Local Research Ethics Committee approved and helped establish the cohort at recruitment and subsequently at each assessment. This study enrolled 1,509 women who gave birth to 1,536 children on the Isle of Wight during the recruitment period, between 1st January 1989 and 28th February 1990. Of these, informed consent was obtained for 1,456 children.

The study children were followed up on six different occasions over the course of 26 years, that is, at one, two, four, 10, 18 and 26 years; study children who participated or not in each follow-up study were compared for non-response bias. Self-completed or telephone questionnaires assessing information on asthma, allergic diseases, respiratory, nasal as well as dermatological symptoms and environmental exposures were completed by most study children or their parents at each clinical assessment. Mothers completed a questionnaire soon after the birth of her child, and at one, two and four years; children completed the questionnaire on their own from the age of 10 years. All participants were seen by a doctor or nurse for an assessment at four, 10, 18 and 26 years; these visits included a physical examination and skin prick tests with inhalant allergens and food allergens. The physical examination included height, weight, and signs of allergic diseases, such as wheeze and eczema. At ages 10, 18 and 26 years, spirometry and bronchial provocation tests were carried out, and blood and urine samples were collected. Lung function tests included FEV_1, FVC and PEF. In addition, children with past or current wheezing were invited to perform a methacholine bronchial challenge. Blood samples were collected for measurement of total serum IgE and for IgE antibodies to common inhalant allergens at birth and then at ages 10 and 18 years. Blood samples were collected at age 18 years for 1,211 participants for genetic and epigenetic investigations.

4.1.1.4 MAAS

The Manchester Asthma and Allergy Study (MAAS) (162) study was designed to investigate the relationship between genetic and indoor environmental factors in the development of atopy, asthma and other allergic diseases. The study area comprised 50 square miles of urban and rural areas of South Manchester and Cheshire, United Kingdom, and the hospitals have a total of 8,000 deliveries per annum. The study recruited women who reported a history of allergy and were in their 8^{th} to 10^{th} week of pregnancy while attending antenatal visits in the area of Wythenshawe and at Stepping Hill Hospitals between October 1995 and July 1997. In total, MAAS recruited 1,184 infants who were followed up at one, three, five, eight, 11 and 16 years of age. Questionnaires used to collect demographic data, history of atopic disorder, respiratory symptoms, pet ownership and environmental exposures were completed by study children or their parents at each clinical assessment.

Parents completed questionnaires soon after the birth of their child, and at one, two, three, five and eight years of age, and questionnaires at the subsequent assessments were completed by the study participants. Further information on the environment, such as dust samples from the parental bed, neonate's bed, bedroom floor, nursery floor, living room floor, and upholstered furniture were collected immediately after birth and then at six months and one year; these samples were analysed for allergen and endotoxin content. At ages five, eight and 11 years, lung function tests, including FEV_1 and FVC, were performed. At age five years, airway reactivity was assessed by means of a eucapnic voluntary hyperventilation challenge, and at ages eight and 11 years, study participants also performed a methacholine bronchial challenge. Atopic sensitization was assessed at five years by skin prick testing with most common inhalants and food allergens. Allergen-specific IgE, IgG and IgG4 were measured at age five years. Biological samples, including serum and plasma, were stored over several time points, and whole-genome genotyping using the Illumina 610 QUAD chip was completed on 935 children.

4.1.1.5 SEATON

The Study of Eczema and Asthma To Observe the effects of Nutrition (SEATON) (161) was established to investigate the effect of maternal diet during pregnancy on the development of asthma and other allergic conditions, such as eczema and hayfever, in her child. This study recruited 2,000 pregnant women during their 12th week of gestation while attending their first antenatal clinical appointment at Aberdeen Maternity Hospital between 1997 and 1999. The Aberdeen Maternity Hospital provides obstetric care for the population of the Grampian region. A total of 5,490 pregnant women enroled. Approval for the study was obtained from the Grampian Research Ethics Committee, and all mothers provided written informed consent. This study included a total of 1,924 children who were followed up at six months and at one, two, five, 10 and 15 years by postal questionnaire; parents completed the questionnaire soon after birth of their child and until the study child was 10 years old. The questionnaires were used to obtain information on sociodemographic characteristics; respiratory symptoms, such as wheeze, breathlessness, rhinitis, doctor's diagnosed asthma, eczema or hayfever; history of parental allergies; number of other children in the house; household pets and infant antibiotic use.

At the ages of five, 10 and 15 years, study children were invited to the hospital for a clinical assessment that included spirometry, skin prick testing to common allergens (cat, timothy grass, egg and house dust mite), serum-specific IgE, measurement of exhaled nitric oxide and methacholine bronchial challenge Spirometry measures included FEV_1, FVC and PEF. Biological samples for DNA extraction were obtained for 626 children.

Table 4.1: Characteristics of the STELAR cohorts

Study name	City/County	Recruitment period	Sample size	Study child's age at last follow-up (years)
Avon Longitudinal Study of Parents and Children (ALSPAC)	Bristol	1991–1992	14,451	24
Ashford	Ashford	1992–1993	642	20
Isle of Wight (IoW)	Isle of Wight	1989–1990	1,456	26
Manchester Asthma and Allergy Study (MAAS)	Manchester	1995–1997	1,184	16
Study of Eczema and Asthma To Observe the effects of Nutrition (SEATON)	Aberdeen	1997–1999	1,924	15

4.1.2 Asthma e-Lab

To maximise the efficiency of the collaboration, STELAR created an innovative scientific online platform (Asthma e-Lab; www.asthmaelab.org) to provide support for description, storage and harmonisation of information from the five UK asthma birth cohorts and to share computational and statistical methods, analytic codes and manuscripts across them (116). The STELAR cohorts were required to share/upload questionnaires, research protocols, consent forms and standard operating procedures together with ethical and legal documents to the Asthma e-Lab coordinating group. As a result, a collaborative dialogue between a broad range of researchers, including clinicians, statisticians, computer scientists and geneticists, was developed through this web-based platform. Characteristics of participating cohorts were documented and stored in the Asthma e-Lab; these characteristics included general study design, number of participants, participant characteristics, recruitment methods and number and type of biological samples collected. Cataloguing such information helped reveal the not only sample sizes available for specific analyses but also the level of heterogeneity across study designs.

The Asthma e-Lab is available via the internet, but access to its functions and stored information is strictly controlled and depends on the user's level of permission, which reflects the tasks the user is expected to carry out. For example, an administrative staff member is unlikely to need to export information from the e-Lab so permission to do that is not given. Authenticated researchers are able to view information and download datasets, but they cannot change information stored in the e-Lab unless they are also a representative of a study who is authorised to carry out data management tasks. The authorised and named data managers are allowed to edit annotations only for the study they represent but not for other cohorts, and they cannot make changes to the e-Lab platform. A request for cohort datasets must come from a researcher who submits an initial proposal. The proposal should have clearly stated aims and/or hypotheses and describe the relevant exposure, outcome and confounders that will be considered, in order to justify access to the datasets that the researcher wishes to use. The proposal is then reviewed by the STELAR Principal Investigators to ensure that it is sound from both an ethical and a scientific perspective.

Once this is approved and signed off by the STELAR team, the researcher is given password-protected access to the resource. It is the responsibility of the researcher to adhere to the specified confidentiality rules at all times, including no attempt to identify individuals or households or to share confidential information with other researchers without permission from STELAR Principal Investigators. Users are also required to notify the STELAR team of any errors discovered in the datasets.

4.2 Data annotation

To facilitate integration of datasets across multiple cohorts, data managers from all five UK cohorts provided explicit information about the structure of their dataset when they uploaded the dataset into the Asthma e-Lab. For the Ashford study conducted as part of my PhD research, I catalogued information collected over 14 years and annotated more than 1,500 variables. This mapping necessitated a complex and detailed description file ('metadata') that had to conform to a highly specific format and be syntactically correct. A number of metadata files were created to describe the meaning of each variable, such as its definition, collection time, coding and data type. The body of the description file contained 18 columns in which several pieces of metadata about each variable were allocated. Most of these columns were compulsory, some were optional. They included

- a 'ColumnIdentifier' which gave a name to the variable column being described and was compulsory;
- a 'ColumnTags' field that was used to tag variables with important words and phrases to facilitate effective searching of variables in the Asthma e-lab;
- a 'Format' column used to describe the representation of variables values when necessary;
- a 'Unit' column that provided a unit of measure for the variable column (entries were restricted to the Unified Code for Units of Measure (UCUM) units);
- 'Input Datatype' and 'Output Datatype' were used to specify which input and output type of values a variable had and what type of mathematical, relational or logical operations could be performed on it without causing an error during the programming process;

- a ‘DataLogic’ descriptor that characterized the variable being described as date, counted number of units, numeric value determined by measuring object or an information establishing the identity of a unique object;
- a ‘ColumnDescription’ field that was used to capture the real-world meaning of the variable in a computer readable format by specific templates (e.g. whether it corresponded to a skin prick test, a measurement of lung function or a response to a questionnaire item); for example, a ‘QuestionResponse’ template was available for responses to questionnaire items and a ‘SPTResult’ template for the results of skin prick tests. For each of these templates, additional information was required, such as a full description of the question or measure collected, specification of the identification variable, respondent of the question, method of follow-up and child’s age at the time of collection;
- ‘DataCodes’ and ‘MetaCodes’ fields that were used to map coded variables as well as their categories.

Once a live dataset with its description file was prepared, this was quality assured and checked for syntax errors through an online validation process and then incorporated into the Asthma e-Lab. **Table 4.2** shows part of a metadata file that I created to annotate and map some of the Ashford variables collected using paper questionnaires since the child’s birth.

Table 4.2: Metadata file created to annotate some of the variables collected using an Ashford questionnaire at age two years

ColumnIdentifier	ColumnDescription	Format	Unit	Input Datatype	OutputDatatype	DataLog
"BIRTH"		mm/dd/yy		Date	Date	Date
"BREAST"	QuestionResponse(QuSubject=Topic.StudySubject|VariableLabel="Did you ever breast feed?"|SubjectNo="Exposuredetails":"STUDYNO"|QuName="ExposureDetails"|MethodOfFollowUp=FollowUpMethod.HomeVisitResidence|Respondent=Person.MotherOfIndexChild|FollowUp=FollowUp.ADExposure8Weeks|AIW="Exposuredetails":"AIW")			String	Term	DomainVe
"HOUSPEOP"	QuestionResponse(VariableLabel="No. of people living in house"|Respondent=Person.MotherOfIndexChild|QuSubject=Topic.ChildHomeEnvironment|SubjectNo="Exposuredetails":"STUDYNO"|QuName="ExposureDetails"|MethodOfFollowUp=FollowUpMethod.HomeVisitResidence|FollowUp=FollowUp.ADExposure8Weeks|AIW="Exposuredetails":"AIW")			int	int	count
"HOUSROAD"	QuestionResponse(VariableLabel="Distance from public road"|Respondent=Person.MotherOfIndexChild|QuSubject=Topic.ChildHomeEnvironment|SubjectNo="Exposuredetails":"STUDYNO"|QuName="ExposureDetails"|MethodOfFollowUp=FollowUpMethod.HomeVisitResidence|FollowUp=FollowUp.ADExposure8Weeks|AIW="Exposuredetails":"AIW")		"m"	int	int	measure
"MUMSMOK"	QuestionResponse(VariableLabel="How many per day (mother)?"|SubjectNo="Exposuredetails":"STUDYNO"|MethodOfFollowUp=FollowUpMethod.HomeVisitResidence|Respondent=Person.MotherOfIndexChild|QuName="ExposureDetails"|QuSubject=Topic.ChildHomeEnvironment|FollowUp=FollowUp.ADExposure8Weeks|AIW="Exposuredetails":"AIW")			int	int	count
"PARTSMOK"	QuestionResponse(VariableLabel="How many per day (partner)?"|SubjectNo="Exposuredetails":"STUDYNO"|MethodOfFollowUp=FollowUpMethod.HomeVisitResidence|Respondent=Person.MotherOfIndexChild|QuName="ExposureDetails"|QuSubject=Topic.ChildHomeEnvironment|FollowUp=FollowUp.ADExposure8Weeks|AIW="Exposuredetails":"AIW")			int	int	count
"SEX"				String	Term	DomainVe
"STUDYNO"	SubjectIdentifier(SubjectGender="Exposuredetails":"SEX"|SubjectDob="Exposuredetails":"BIRTH")			string	string	identifie

4.3 Data availability and harmonisation among the STELAR cohorts

Data harmonisation is the process of combining/integrating information from different sources, potentially having different formats and naming conventions, and transforming it into a single cohesive dataset for analysis. This procedure determines the final form of the dataset and ensures a bigger sample size, better use of collected information and enhances the quality of the work throughout application of effective statistical analysis as well as research collaboration. If datasets are not harmonized, each source must be explored separately, which can lead to loss of information and poor generalizability of results.

The harmonization process is therefore a key procedure and can be described by three phases—identification, transformation and matching of variables. In the first phase, different sources are screened and examined in order to search for, identify and map comparable variables collected at similar time points. In the second step, variables are transformed via cleaning and format conversions to create homogenous information among sources and transform it into a common structure. This phase also involves detecting, correcting or removing inaccurate or erroneous values and recoding variables. This phase can be the most challenging and laborious part of data harmonization, because each source comes with its own set of unique issues, and an error can cause inaccurate or missing values throughout. Third, the common key variables and features of sources are matched, and the harmonised dataset obtained can be used for statistical analysis.

I harmonized similar variables across the STELAR cohorts in order to standardize, compare as well as pool similar information over a period of time. Details of their harmonization are described in the following three sections.

4.3.1 Identification of variables among the STELAR cohorts

During the design of the Ashford survey at age 20 years, available STELAR questionnaires were reviewed to identify important variables for my research aims and to assess information availability among STELAR cohorts.

The Asthma e-Lab was then used to efficiently search for and download annotated variables by using Boolean operators such as 'OR' and 'AND' as well as 'wildcards'. The laborious procedure of searching and downloading variables from the e-Lab is described below, and figures are shown as representations of instructions.

1. The user clicks on 'Cohort Manager' to enter in the search webpage (**Figure 4.1**)

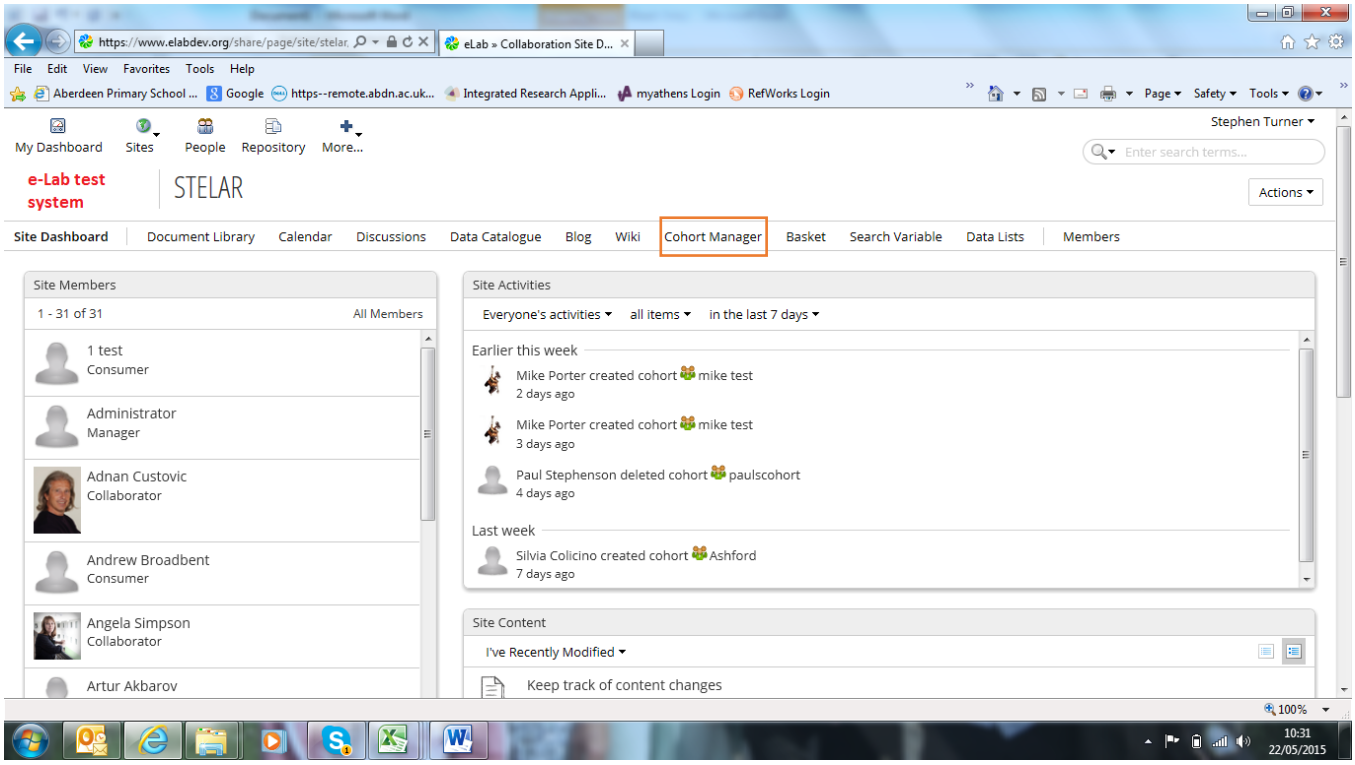


Figure 4.1: Asthma e-Lab home page

2. The user selects the cohort(s) from which s/he wants to download information then searches for variables using the 'Search Variable' button and types the term that s/he is interested in. Below I show the result from searching on 'smoking'; note that only the cohort(s) that the user is interested in remain 'green ticked' (**Figure 4.2**).

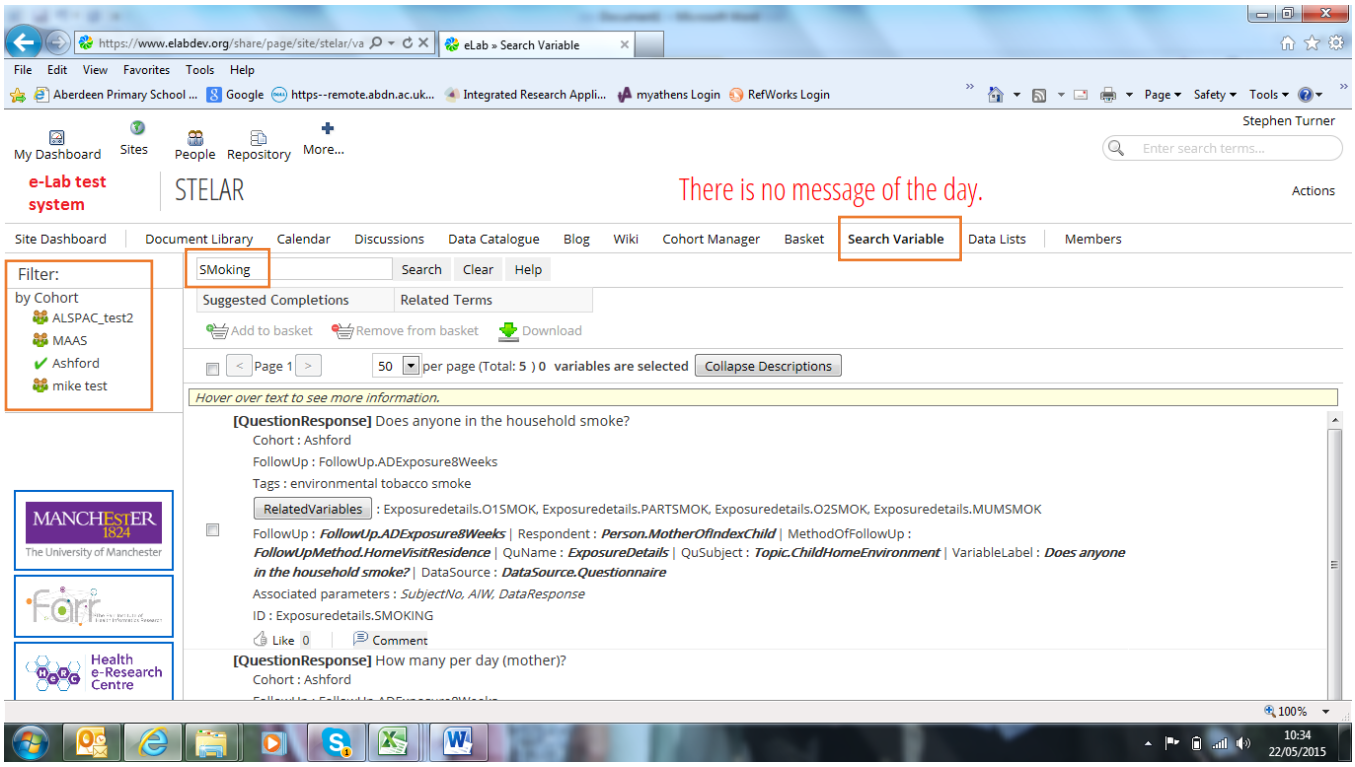


Figure 4.2: Results from searching on 'smoking' in the Ashford cohort

3. The user ticks the boxes of variable(s) that s/he wants to download and then clicks 'Add to Basket' in order to store them in the online basket (**Figure 4.3**). The user is allowed to edit his/her choice of variables whilst they are stored in the basket.

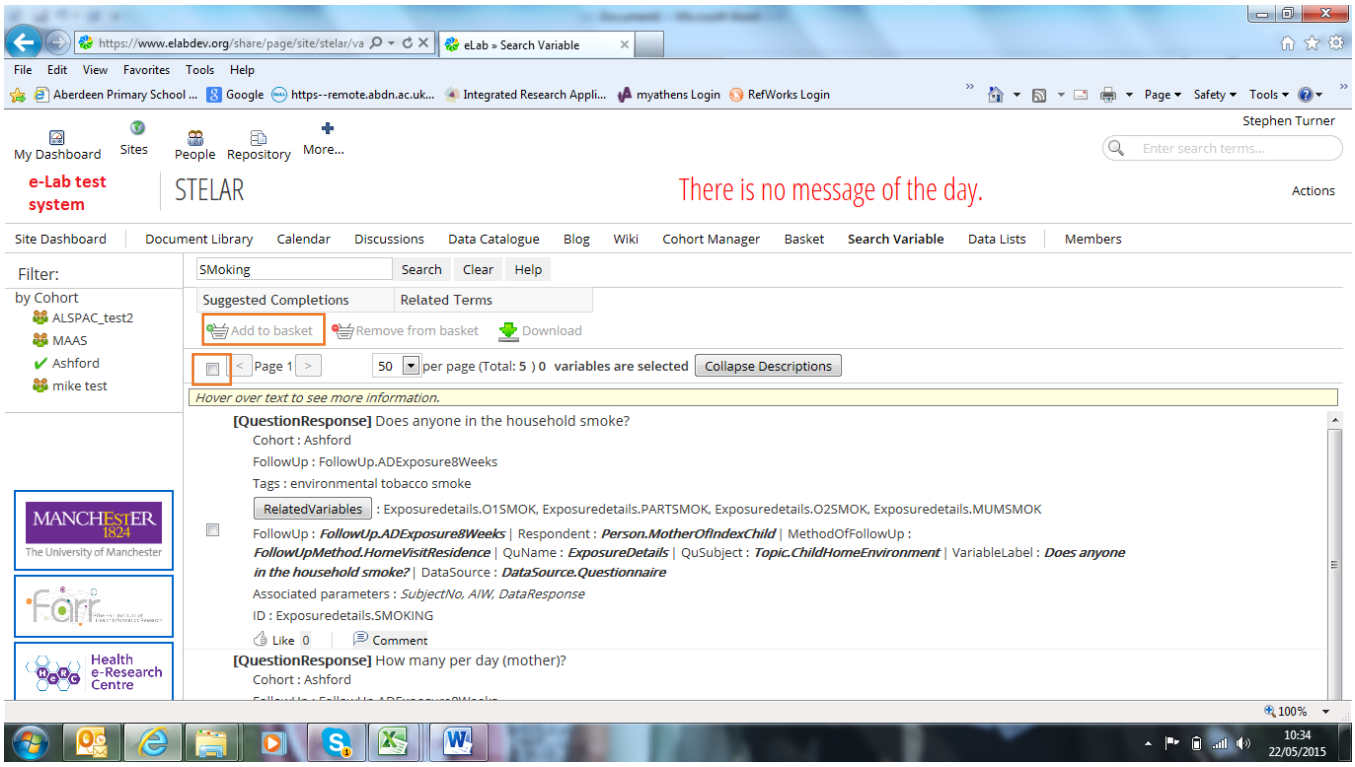


Figure 4.3: Selection of variables of interest and inclusion in the online basket

Overall, I identified and mapped 69 variables that were available across all five cohorts; 41 of those variables were considered comparable and compatible among the STELAR cohorts as they were collected at similar time points and used similar wording for the questions (**Table 4.3**). Data for those variables were downloaded from the Asthma e-Lab. They included demographic characteristics, parental history of allergic diseases, child's medical information and relevant respiratory symptoms from birth to age 10 years as well as asthma-related information in adolescence, that is, between ages 15 and 20 years.

Table 4.3: Comparable variables among the STELAR cohorts downloaded from the e-Lab

Grouping	Variable	Collection time	ALSPAC	Ashford	IoW	MAAS	SEATON
Demographic	Study cohort	Recruitment	✔	✔	✔	✔	✔
	Sex	Birth	✔	✔	✔	✔	✔
	SES (based on parental occupation)	Recruitment	✔	✔	✔	✔	✔
Parental allergies	Maternal asthma	Recruitment	✔	✔	✔	✔	✔
	Maternal eczema	Recruitment	✔	✔	✔	✔	✔
	Maternal hayfever	Recruitment	✔	✔	✔	✔	✔
	Paternal asthma	Recruitment	✔	✔	✔	✔	✔
	Paternal eczema	Recruitment	✔	✔	✔	✔	✔
Postnatal exposures	Older siblings	By 3 yrs	✔	✔	✔	✔	✔
	No. of older siblings	By 3 yrs	✔	✔	✔	✔	✔
	Birth order	By 3 yrs	✔	✔	✔	✔	✔
	Maternal age	Recruitment	✔	✔	✔	✔	✔
	Premature birth	Birth	✔	✔	✔	✔	✔
	Ever breast fed	By 3 yrs	✔	✔	✔	✔	✔
	Vaginal delivery	Birth	✔	✔	✔	✔	✔
	Age at weaning (weeks)	By 3 yrs	✔	✔	✔	✔	✔
Environmental exposures	Cat ownership	1 yr	✔	✔	✔	✔	✔
		4 or 5 yrs	✔	✔	✔	✔	✔
	Dog ownership	1 yr	✔	✔	✔	✔	✔
		4 or 5 yrs	✔	✔	✔	✔	✔
	Moved house	By 2 yrs	✔	✔	✔	✔	✔
	Maternal smoking	Recruitment	✔	✔	✔	✔	✔
		1 yr	✔	✔	✔	✔	✔

Abbreviations - SES: socioeconomic status; yrs: years; DD: doctor's diagnosis; SPT: skin prick test; HDM: house dust mite.

Table 4.3: Comparable variables among the STELAR cohorts downloaded from the e-Lab *(continued)*

Grouping	Variable	Collection time	ALSPAC	Ashford	IoW	MAAS	SEATON
Symptoms	Current wheeze	1 yr	✔	✔	✔	✔	✔
		2 or 3 yrs	✔	✔	✔	✔	✔
		4 or 5 yrs	✔	✔	✔	✔	✔
		8 or 10 yrs	✔	✔	✔	✔	✔
	Frequency of wheezing in the past 12 months	1 yr	✔	✔	✔	✔	✔
		2 or 3 yrs	✔	✔	✔	✔	✔
		4 or 5 yrs	✔	✔	✔	✔	✔
		8 or 10 yrs	✔	✔	✔	✔	✔
	Eczema in the past 12 months (DD and parent reported)	1 yr	✔	✔	✔	✔	✔
		2 or 3 yrs	✔	✔	✔	✔	✔
		4 or 5 yrs	✔	✔	✔	✔	✔
	Ever DD asthma	4 or 5 or 7.5 yrs	✔	✔	✔	✔	✔
	Asthma treatment in the past 12 months	4 or 5 or 6.4 yrs	✔	✔	✔	✔	✔
Skin prick tests	SPT (cat/HDM/mixed grass)	4 or 5 or 7 yrs	✔	✔	✔	✔	✔
Asthma-related information in young adulthood	Current wheeze	15 yrs					✔
		16 yrs	✔			✔	
		18 yrs			✔		
		20 yrs		✔			
	Asthma treatment in the past 12 months	15 yrs					✔
		16 yrs	✔			✔	
		18 yrs			✔		
		20 yrs		✔			

Abbreviations - SES: socioeconomic status; yrs: years; DD: doctor's diagnosis; SPT: skin prick test; HDM: house dust mite.

4.3.2 Transformation of the STELAR variables

After I identified variables of interest, I applied a transformation step to produce homogeneous information across the STELAR cohorts. The transformation process, performed using scripts created in R software (https://www.r-project.org), included importing the data, recoding variables (e.g. changing '2' = no to '0' = no), replacing missing values due to skipped questions with appropriate values and deriving new variables by merging different data types to obtain 'compromises' among STELAR cohorts.

For example, a data compromise concerns frequency of wheezing. The study questionnaires asked several questions about asthma-like symptoms at each follow-up, such as the number and severity of wheezing episodes in the past 12 months. However, this question was asked slightly differently among the STELAR cohorts; some of them sought the actual number of wheezing episodes, but others sought free text or check-box answers. Hence, I had to transform continuous variables into categorical variables and reassign values in order to have similar probability distribution and frequency of categories among the cohorts. An analogous data compromise was performed for socioeconomic status (SES), which was calculated based either on parental occupation at recruitment using the Standard Occupational Classification (SOC) or the National Statistics Socio-Economic Classification (NS-SEC). SOC is a commonly used classification of occupations in the UK, while the NS-SEC, which measures the employment relations and conditions of occupations, was revised based on the SOC. Since the STELAR cohorts used slightly different socioeconomic classification, I derived the final five socioeconomic classes among the STELAR cohorts as follow: 1) managerial, technical, administrative and professional occupations at large employers, higher and lower managerial, professional, higher technical and administrative occupations; 2) skilled non-manual occupations and intermediate occupations, such as clerical sales and service; 3) skilled manual occupations, including small employers and own-account workers; 4) partly-skilled occupations, including lower supervisory and lower technical occupations; and 5) semi-routine and routine occupations.

Another example of compromise pertains to skin prick test results. The test was considered positive if there was a wheal 3 *mm* or greater than that of the saline control to any of the three main allergens (cat, mixed grass and *Dermatophagoides pteronyssinus* [house dust mite]). Skin prick testing was performed at a similar age, around five years old, among all cohorts except the ALSPAC cohort (seven years). Similarly, during middle childhood, information on whether asthma was ever diagnosed by a doctor and the use of asthma medication in the past 12 months was collected at different ages among the STELAR cohorts. The ALSPAC study reported this information at 77 and 91 months, and in the other cohorts it was collected around age five years (60 months).

Further compromises include current wheeze, which was assessed in each follow-up questionnaire by similar questions among the STELAR cohorts and was defined as a positive response to the following question "Has your child had wheezing with whistling in the last 12 months?" in the ALSPAC study and in the other four cohorts, to the question "Has your child had wheezing *or* whistling *in the chest* in the last 12 months?". Information on eczema in the past 12 months was derived based on parent-reported information and previous doctor's diagnosis. ALSPAC collected this information at 18 months, 30 months and 57 months, but the other cohorts assessed eczema at one, two or three, and five years of age. Similarly, questions about colds or runny nose symptoms in the past 12 months were asked at 18 months in ALSPAC and at one year of age in the remaining cohorts.

Parental allergies (maternal asthma, eczema and hayfever and paternal asthma and eczema) were assessed at recruitment by self-report or reported doctor's diagnosis among the STELAR cohorts. Maternal age at birth of the study child either was recorded in the questionnaire or was derived based on maternal date of birth and date of birth of the study child. Information on pet ownership, specifically, cats and dogs, during the first year of the study child's life and around his/her fourth or fifth year was assessed by a question about the presence of pets in the house among all STELAR cohorts except ALSPAC, which asked about pet contact instead. Questions about plans to move to a new home or the number of times the child moved were asked at the birth of the study child in the Ashford and MAAS studies, around age one year in SEATON and IoW cohorts and around two years of age (21 months) in ALSPAC. Birth order (the chronological order of sibling births in a family) was derived based on the number of older siblings; I classified this information by using three ordinal

categories— first born, second born and any other birth-order position. Premature birth was defined as a birth that took place more than three weeks before the baby's estimated due date and thus occurred before the 37th week of pregnancy. Information such as ever breastfed, referring to infants who had been put to the breast at least once, and age at weaning, indicating the age, in weeks, at which infants started eating solid food, was collected and recorded up to two or three years of age among the STELAR cohorts. Therefore, I derived two variables on breastfeeding and age at weaning based on data collected at any time during the first two or three years.

4.3.3 Matching of variables among the STELAR cohorts

To create a single, harmonised dataset, the transformed variables from the STELAR cohorts were matched and merged by type of data and by specific time points. The dataset was assessed for outliers, missing values, duplicate records, and distribution of the data before using it for the development of the prediction models reported in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years) and Chapter 6 (Further investigations using the Bayesian model). It is worth noting that matching and harmonization of information gathered in adolescence was established at slightly different time points among STELAR cohorts; ALSPAC collected this information at 16.6 years, Ashford at 20 years, IoW at 18 years, MAAS at 16 years, and SEATON at 15 years. Asthma definition and statistical analysis on asthma prediction in adolescence will be described in detail in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years) and Chapter 6 (Further investigations using the Bayesian model).

4.4 Conclusion

Data harmonisation identifies the procedure used to standardize, compare as well as pool similar information among different sources over a period of time. This procedure allowed me to maximize the use of the STELAR data collected over 20 years producing a bigger sample size and, in turn, empowering the efficiency of statistical analysis. In this chapter, I have described the five STELAR asthma UK birth cohorts and how I approached the data harmonisation procedure within them in order to create my final dataset. The process was lengthy and challenging due to discrepancies in variable recording between different cohorts such as different names, categorizations and units of measurement. I identified and harmonised 41 variables; 39 of these variables will be used as

candidate predictors, while the other 2 variables will be used to define the primary outcomes in the main statistical analyses of asthma prediction in adolescence described in detail in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years) and Chapter 6 (Further investigations using the Bayesian model).

Chapter 5 – Comparison of statistical methods for the prediction of asthma at 15–20 years

Predictive models are statistical methods used in medicine to assist clinicians during the decision-making process as they attempt to determine a future clinical outcome (event) based on an individual's risk profile. These models may help with diagnosis and prognosis of the disease of interest by predicting the probability that a subject with a given set of predictors will develop the disease and by identifying subjects at high risk of developing a given disease later in life (120, 123). Therefore, the fundamental objective of predictive modelling is to find an optimal and parsimonious predictive model that provides good predictive performance with as few predictors as possible.

According to established guidelines (101, 104, 105, 118-123, 125), prediction models should be developed using the following five-phase process: 1) definition of target population, 2) definition of clinical outcome, 3) selection of candidate predictors, 4) model development and 5) validation of predictive performance. With regard to the first two stages—defining a clear and appropriate definition of clinical outcome—the age of the population used for outcome prediction as well as the age of the target population are necessary to develop an accurate and useful clinical predictive model. The definitions of these aspects affect database structure, determine the specification of the predictive model and influence predictive performance. With regard to the third stage, evaluation and selection of the most relevant predictors are essential; in general, inclusion of more than 10 predictors in the final model can decrease its simplicity, interpretability and convenience (102).

Predictors found to be significantly associated with the clinical outcome in previous research should be tested as candidate predictors.

With regard to the fourth stage, a variety of statistical methods can be used to develop predictive models, such as standard regression analysis, Machine Learning and Bayesian methods, which include a number of candidate predictors based on subjects' characteristics, environmental exposures and clinical features (196-199). These candidate predictors are factors that are thought to be associated with the health outcome of interest and are usually selected on the basis of clinical experience and literature reviews. In practice, the number of events in the datasets that are used to develop the predictive model is often too low relative to the number of candidate predictors, particularly when the event of interest is relatively rare and a large number of predictors possible. In such situations, predictive models may be too complex and lead to 'model overfitting' and poor predictive performance (105, 113, 118, 200). An overfitted model often accurately predict events for subjects in the dataset used to develop the model, but performs less well for a new group of subjects. To overcome this issue, variable selection techniques such as regularisation methods have been suggested.

For the fifth stage, the predictive performance of the model is evaluated in data that have structural and population-level characteristics similar to the data from which the model was developed (99, 105, 110). Evaluation of predictive performance can be performed using internal or external validation. To ensure the generalisability of a predictive model, it is necessary to evaluate its performance empirically in independent validation studies (external validation) with subjects from other, but similar, populations that were not used to develop the model (99, 100, 105). This process assesses the general applicability of a predictive model to an independent sample and reduces model overfitting (201). The predictive performance of a model is measured by its calibration and discrimination measures (114, 202). Calibration measures refer to the agreement between observed and predicted values and are usually assessed graphically and/or using the Hosmer-Lemeshow goodness-of-fit test. Discrimination measures refer to the ability of a model to distinguish between subjects with and without the disease of interest; these measures include sensitivity, specificity, positive and negative predictive values and area under the ROC curve.

The main aim of this chapter is to describe the application and evaluation of different statistical methods to the predict asthma in individuals between age 15 and 20 years by using information gathered about them in early childhood (before age five years), from the asthma UK birth cohorts included in STELAR. I will evaluate both classical and modern techniques for performing model selection and for predicting an individual's risk of developing asthma in adolescence. In particular, I describe regression techniques, from a simple step-wise regression to least absolute shrinkage and selection operator (LASSO) and elastic net (EN), as well as empirical Bayes and Bayesian regularization methods. I discuss the strengths as well as limitations of such methods. The predictive models developed will be 1) externally validated to assess their generalisability and the transportability of the results and 2) evaluated and compared through measures of accuracy, discrimination and calibration.

5.1 Methods

5.1.1 Definition of the target population

The analysis presented in this chapter was performed in a subgroup of STELAR children whose parents reported that they had wheezing episode(s) between the ages of two and five years. Such children are at higher risk of developing asthma later in life and thus might benefit from early diagnosis of asthma and therapeutic management (98). This population is of clinical interest as they are more likely to have persistent asthma in adolescence; henceforth, this population will be defined as a 'high-risk population'.

This group involves 1,008; 145; 230; 115 and 178 children from the ALSPAC, Ashford, IoW, MAAS and SEATON studies, respectively (N = 1,676). Details on the STELAR cohorts can be found in Chapter 4 (Data sources, paragraph 4.1.1).

5.1.2 Definition of outcome

Current asthma was defined as the outcome and was assessed between 15 and 20 years of age, in order to predict the condition in adolescence. I defined current asthma by a positive answer to each of the following items of the questionnaire: wheezing in the past 12 months and asthma medications in the past 12 months.

Asthma medications include relievers (short-acting bronchodilators), long-acting bronchodilators and inhaled or oral corticosteroids. Information on current wheeze and current asthma medications was collected at comparable ages among the five UK asthma birth cohorts (for details, see Chapter 4 - Data sources, paragraph 4.3)**,** specifically, at 16.6 years in ALSPAC, at 20 years in Ashford, at 18 years in IoW, at 16 years in MAAS, and at 15 years in SEATON.

5.1.3 Potential predictive variables

Initially, I reviewed the literature to identify factors potentially associated with the occurrence and persistence of asthma in later childhood and adolescence (85, 86, 141, 203). I then harmonised the related variables across the five STELAR cohorts. Details on the harmonisation procedure as well as a list of candidate predictors can be found in Chapter 4 (Data sources, paragraph 4.3). Although I harmonized 39 candidate predictors among the STELAR cohorts, in the analyses presented in this chapter I only considered information collected in early childhood (by age five years), as this time frame is in line with existing asthma-predictive models and available systematic reviews (85, 86, 141). In total, I included 32 candidate predictors in my predictive models. These data were collected at recruitment and then from birth to age five years across all five cohorts, and they consisted of demographic characteristics, parental allergies, perinatal data, environmental exposures and respiratory and eczema symptoms.

5.2 Statistical analysis

Statistical analysis was performed mainly using a two-step approach. In the first step, referred to as the model development process, I identified important predictors associated with the outcome and estimated their effects by using several approaches, including stepwise logistic regression and different regularization logistic models within classical, Bayesian and empirical Bayes frameworks. The second step consisted of evaluating the performance of the predictive models generated in the first step by using measures such as sensitivity, specificity, area under the ROC curve and predictive values. Model performance was evaluated in external populations in order to appraise the reliability of each predictive model in independent studies. Results reported in this chapter come from a multiple imputation analysis. The Transparent Reporting of a multivariable predictive model for Individual Prognosis Or Diagnosis (TRIPOD) guideline (125) was adopted to describe the work and

report the main findings of my predictive models and their performance measures and to enable a fair comparison with existing childhood asthma-predictive models.

The ALSPAC and Ashford cohorts had similar inclusion criteria and prevalence of asthma, whilst these were different in the IoW, MAAS and SEATON cohorts. For this reason, the ALSPAC and Ashford cohorts (development/training data) were used to develop the predictive models, and the other studies (validation/testing data) were employed to externally validate their predictive abilities. Data were prepared and analysed using R software, version 3.3.3.

5.2.1 Model Development

Univariable (crude) associations between candidate predictors and asthma at the age of 15–20 years were assessed in the ALSPAC and Ashford studies. Differences were analysed using the chi-square test for categorical data and Student *t*-test for normally distributed continuous data.

I applied the following five statistical techniques to develop the predictive models: stepwise logistic regression; two regularization logistic models developed in a frequentist framework—least absolute shrinkage and selection operator (LASSO) and elastic net (EN); a Bayesian and an empirical Bayes regularization method. All these approaches enable the selection of important predictors and the simultaneous estimation of their regression coefficients. The regularization regression methods address the issues of overfitting and multicollinearity that occur when two or more predictor variables are highly correlated with one another, that is, when one variable can predict the other variable (113, 118). Specifically, these regularization methods impose constraints on the values of the regression coefficients and the size (number of predictors) of the models; these constraints are characterised by shrinkage parameters that pull the regression coefficients toward zero, allowing variable selection. In the analyses, I dichotomised all categorical variables with more than two categories into dummy variables before including them into the models. The dichotomisation process is necessary to perform the regularization regression methods using the package '*glmnet'* in the statistical software R. Thus, I derived 46 variables from the initial 32 candidate predictors.

5.2.1.1 Classical stepwise regression

Stepwise logistic regression, which is the commonest method used to perform variable selection and develop predictive models in medical research, assesses the strength of the association between a set of risk factors or predictors and a binary outcome of interest (204, 205). Logistic regression can be difficult to apply in high-dimensional data where the number of predictors, Q, is large relative to the sample size, N; in this context, overfitting and multicollinearity become important issues as they lead to unstable estimates and poor predictive performance. To avoid potential overfitting and improve predictive performance, it is necessary to select an optimal set of predictors $\boldsymbol{X} = (X_1, X_2, \dots, X_Q)$ for inclusion in the predictive model (113, 118). Several variable selection methods have been proposed for multiple logistic regression, such as backward elimination, forward selection and stepwise selection (204, 205). I first briefly describe the logistic regression method and then explore the stepwise selection method.

In logistic regression the dependent variable, $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_N)$, comes from a Bernoulli distribution and assumes two values, 0 and 1, where 1 indicates the presence of the condition, such as asthma in adolescence, given a set of predictors (206).

$$Y_i | X_{1i}, \dots, X_{Qi} \sim Bernoulli\ (p_i) \qquad i = 1,2, \dots, N$$

where Y_i and X_i are the outcome and set of predictors, respectively, of the *i-th* subject; N is the number of observations and Q is the number of candidate predictors.

The probability of the condition of interest occurring for a given individual and a set of predictors $P(Y_i = 1 | X_{1i}, \dots, X_{Qi}) = p_i$ can be calculated using the following formula:

$$p_i = \frac{e^{\beta_0 + X_i^T \boldsymbol{\beta}}}{1 + e^{\beta_0 + X_i^T \boldsymbol{\beta}}} \qquad i = 1,2, \dots, N$$

where β_0 is the estimate of intercept and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_Q)$ is the set of regression coefficients used to estimate the probability of the condition to occur. Typically, these regression coefficients are calculated using maximum likelihood estimation (206). The likelihood estimate measures the data's evidence based on the unknown regression coefficients:

$$L(\beta_0, \boldsymbol{\beta}) \propto \prod_{i=1}^{N} p_i(1 - p_i)$$

However, for computational convenience, the log-likelihood is often maximised (or the negative log-likelihood function is minimized) to estimate the unknown parameters:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \ argmax \left\{ \sum_{i=1}^{N} log \left(1 + e^{\beta_0 + X_i^T \boldsymbol{\beta}}\right) + \left(Y_i - \ (\beta_0 + \ X_i^T \boldsymbol{\beta})\right)\right\}. \quad [1]$$

The stepwise selection method for multiple regression is the commonest variable selection approach in the medical field (200, 207). At each step, this method tests variables to be included or excluded by using a combination of forward selection and backward elimination; the best subset of candidate predictors is usually searched by the Akaike information criterion (AIC) at each step so that the final model is the model with the lowest AIC value (122, 208, 209). The AIC, which is a measure of the quality of each model for a given set of data and is based on a penalty of maximum log likelihood (113), identifies the best-fitting model that uses the fewest variables. In the R software, the selection of a model based on the AIC value in the stepwise algorithm can be performed using the '*step*' function.

Despite the simplicity of this method, the stepwise technique can give overestimated regression coefficients, *p*-values that do not have the proper meaning as they are not corrected for multiple testing and problems regarding the presence of collinearity (200, 204). Multiple testing involves simultaneous hypothesis testing on regression coefficients, and it can lead to overestimation of the true effect size; in the stepwise procedure, a number of models are established and multiple hypothesis testing is conducted to identify regression coefficients that are significantly different from zero and that can be included into the final model (200, 204). A 95% confidence interval is computed for each regression coefficient; this interval includes the true value of the parameter in 95% of occasions.

5.2.1.2 Classical regularization regression analysis

The classical regularization regression methods are flexible shrinkage approaches that can be performed using the R package '*glmnet'*. These methods fit the same statistical model as logistic regression but use a different estimation procedure by placing a constraint on the values of the regression coefficients (210). This constraint is characterised by hyperparameters (shrinkage parameters) that control the amount of shrinkage, that is, the degree to which the regression coefficients are pooled toward zero, leading to selection of predictors and a sparse model. These shrinkage parameters may be fixed to a constant term or calculated via some criterion, for instance minimizing mean-square error, deviance or mean cross-validated error by cross-validation methods such as the *k*-fold technique (211), or bootstrapping (212). The mean-square error measures the average difference between observed and predicted values. The deviance is two times the log-likelihood ratio of the full model compared to a reduced one. Cross-validated error represents the prediction error associated with the models developed.

Several regularization regression methods that use different constraints have been proposed (213); the LASSO (Least Absolute Selection and Shrinkage Operator) (214) and elastic net (EN) (215) are the most popular methods. The LASSO coefficients ${\beta_0}^{LASSO}$ and $\boldsymbol{\beta}^{LASSO} = (\beta_1, \beta_2, \dots, \beta_P)$ are estimated by minimizing the negative log-likelihood function with the penalty term, $|\boldsymbol{\beta}| = \sum_{j=1}^{Q} |\beta_j|$ (216), which represents a sum of the absolute value of regression coefficients. Thus, equation [1] becomes

$$(\hat{\beta}_0^{\,LASSO}, \widehat{\boldsymbol{\beta}}^{LASSO}) = argmin \left\{ \sum_{i=1}^{N} \left[log\left(1 + e^{\beta_0 + X_i^T \boldsymbol{\beta}}\right) - \left(Y_i - (\beta_0 + X_i^T \boldsymbol{\beta})\right) \right] + \lambda_1 |\boldsymbol{\beta}| \right\}. \quad [2]$$

where Y_i and X_i are the outcome and set of predictors, respectively, of the *i-th* subject; N is the number of observations; Q is the number of predictors and λ_1 is the shrinkage parameter $(\lambda_1 > 0)$ that controls the amount of shrinkage. The logistic LASSO regression coefficient estimates are usually smaller than those obtained by standard logistic stepwise regression (214). One limitation of LASSO is that the number of selected predictors cannot exceed the sample size. The other limitation occurs when there are groups of correlated variables, where it tends to select the most important variable and ignore the others (128).

Elastic net (EN) overcomes those limitations by introducing a quadratic term, $\boldsymbol{\beta}^2 = \sum_{j=1}^{Q} {\beta_j}^2$, to the penalty function (215). This elastic net penalty guarantees the selection and estimation of coefficients of all the predictors that are highly correlated to one another (215). The estimates from EN are defined by

$$(\widehat{\beta_0}^{EN}, \widehat{\boldsymbol{\beta}}^{EN}) = argmin \left\{ \sum_{i=1}^{N} \left[log\left(1 + e^{\beta_0 + X_i^T \boldsymbol{\beta}}\right) - \left(Y_i - (\beta_0 + X_i^T \boldsymbol{\beta})\right) \right] + \lambda_1 |\boldsymbol{\beta}| + \lambda_2 \boldsymbol{\beta}^2 \right\} \quad [3]$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ are the shrinkage parameters $(\boldsymbol{\lambda} > 0)$ that control the amount of shrinkage.

In contrast to the stepwise regression method, these two regularization regression models do not provide standards errors but introduce a bias toward zero in the estimation of coefficients due to the shrinkage (217).

5.2.1.3 Bayesian regularization analysis

In frequentist statistics the estimation process is driven by the observed data, and model assumptions such as linearity, normality of residual as well as homoscedasticity of variances need to be checked to obtain valid inferences. Estimation in the Bayesian approach is performed by combining the information from a prior distribution, which represents the knowledge and uncertainty about the parameter before the current data are examined, with the likelihood of the data to provide a posterior distribution for the parameter (218). This new probability distribution (posterior probability) is then used to perform all inferences, including estimating regression coefficients and computing standard errors. Treating the regression coefficients as fixed or using a probability distribution is the main difference between frequentist and Bayesian methods (218).

The Bayesian regularization method (219) implemented by the R package '*R2BGLiMS*' is useful as it enables one to identify a subset of predictors associated with the outcome and the possibility of incorporating prior information into the analysis. Variable selection is performed by specifying a probabilistic prior over a group of possible models (220), which depends on a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_Q)$ that indicates whether each covariate is included into the predictive model (221). The binary variable γ_j can assume two values, 0 and 1, where 1 indicates the inclusion into the predictive model of the *j-th* predictor with a related regression coefficient different from zero,

$\beta_j \neq 0$. The number of predictors included in the predictive model (the model size) therefore depends on the prior distribution of $\boldsymbol{\gamma}$.

Specifically, I used a Beta-Binomial prior corrected for the number of candidate predictors (222); the prior distribution of $\boldsymbol{\gamma}$ is a Binomial distribution where the probability of including a predictor into the model, p_j, is not fixed but random. This probability, p_j, follows the Beta distribution and is corrected for the number of candidate predictors. This can be specified as follows:

$$\gamma_j \sim Binomial(1, p_j) \qquad j = 1,2, \dots, Q \quad [4]$$

$$p_j = Beta(1, Q)$$

The identified non-zero regression coefficients were then assigned independent Normal prior distributions centred in 0 with unknown variance, $\sigma^2_{\boldsymbol{\beta}}$. This prior distribution represents the uncertainty of the β_j parameters and was combined with the likelihood of the data to provide a posterior distribution and to compute the regression coefficients and their credible intervals.

$$\beta_j | \gamma_j \sim N(0, \sigma^2_{\boldsymbol{\beta}}) \qquad j = 1,2, \dots, Q \quad [5]$$

Finally, I specified a vague Normal prior for the intercept, β_0, with a very large variance to account for the *a priori* uncertainty about the value of the parameter (223)

$$\beta_0 \sim N(0, 10^6) \quad [6]$$

Models such as this are complex and often analytically intractable, as they require the integration of a large number of unknown parameters to calculate the posterior probability. Markov Chain Monte Carlo (MCMC) is a family of algorithms used to draw several independent values from a probability distribution that are then used to perform approximate inference. MCMC starts with a random initial value and, based on it as well as on the distribution of the parameter of interest, a new random value is chosen; each new value is used to generate the next one. This iterative process continues until an MCMC chain converges to a value, which represents the estimate of the parameter of interest (220). To minimise the impact of the choice of initial values on model convergence and therefore results, a number of first iterations is usually discarded by a process called burn-in.

In particular, for every Bayesian model fitted, I ran one million iterations after a burn-in of one million iterations. The convergence and mixing of iterations were checked visually by using autocorrelation plots showing the correlation pattern among iterations and trace plots showing the parameter values across iterations. For a graphical display of the results, I used the Manhattan plot, which compares posterior probabilities among predictors.

I performed sensitivity analyses by changing the parameter of the Beta distribution (that is, Q, the number of candidate predictors) in order to increase or decrease sparsity. The higher the value of Q, the lower the mean and the variance of the Beta distribution and therefore the lower the probability of including a candidate predictor into the model; a value of the latter that approaches zero allows for variable selection. Specifically, I used a flat uniform prior, $p_j = Beta(1, 1)$, as well as more informative priors that specify a set of values for Q, including 9, 19 and 27. For each regression coefficient, I computed the 95% credible interval, which can be described as the Bayesian parallel of the 95% confidence interval. It indicates the range of values in which the true value of the parameter lies, with 95% probability, given the data and the model.

5.2.1.4 Empirical Bayes regularization analysis

The empirical Bayes (EB) selection approach (224, 225) can be described as a compromise between classical and Bayesian regularization methods. This approach attempts to combine the strengths of frequentist and Bayesian statistics. I performed this method using two R packages—'*EBelasticNet.BinomialCV*', which allows for the estimation of the optimal hyperparameters, and '*EBelasticNet.Binomial*', which accomplishes variable selection and estimation of regression coefficients.

The empirical Bayes method starts with a Bayesian model in which a Normal prior distribution is defined for the regression coefficients (226). However, this distribution is specified by other parameters, referred to as hyperparameters (224, 225), that are estimated exclusively through the data, typically using a *k*-fold cross-validation technique as for the frequentist regularization regression methods. These hyperparameters, similar to those used in LASSO and EN, enable variable selection by controlling the degree of shrinkage and the proportion of non-zero coefficients.

For the statistical analysis reported in this chapter, optimal values of the hyperparameters were chosen through 10-fold cross-validation and then used to estimate regression coefficients of the selected predictors. This model employs a three-level hierarchical prior distribution to model the sparsity and to identify a sensible set of predictors associated with the health outcome. At the first level, I assigned a non-informative Uniform prior distribution to β_0, namely, $p(\beta_0) \propto 1$, while the other regression coefficients follow a Normal distribution with mean zero and unknown variance,

$$\beta_j \sim N(0, \sigma_j^2) \qquad j = 1,2, \dots, Q \quad [7]$$

At the second level, the unknown variance follows an independent Exponential distribution depending on a common parameter, δ,

$$\sigma_j^2 \sim Exp(\delta) \quad j = 1,2, \dots, Q \text{ and } \delta > 0 \quad [8]$$

At the third level, this common parameter follows a Gamma prior distribution where the shape, a, and the inverse scale parameter, b, are the hyperparameters obtained by a 10-fold cross-validation technique:

$$\delta \sim \Gamma(a, b) \quad [9]$$

Therefore, this approach estimates the variance, σ_j^2, from the data and then acquires the posterior distribution for the regression coefficients, β_j, based on those estimated variances (225). Empirical Bayes regularization methods have two major advantages. First, the posterior distribution of results provides standard errors together with estimates of coefficients, as opposed to LASSO and EN in the frequentist framework. Second, they do not require specification of the prior distribution of hyperparameters because they are estimated from the data through a validation method (224, 225). However, this empirical Bayes approach has one major drawback: it uses the observed data twice – first to compute the hyperparameters and then to develop the posterior distribution – leading to possible overconfident results and invalid statistical inference.

5.2.2 Model validation

The performance of the predictive models I developed was validated externally in the IoW, MAAS and SEATON cohorts. Specifically, I used the regression coefficients of each predictive model developed to calculate an individual's probability of having asthma in adolescence across the three validation STELAR cohorts and then classified those children as diseased or not, based on their risk. Accordingly, I assessed the predictive ability of the models throughout their overall performance, calibration and discrimination measures (114).

An important overall performance measure is the Brier score (114); it can assume values from zero to one, with smaller values representing better results. This measure is generally deconstructed into three components—uncertainty, reliability and resolution. The uncertainty quantifies the uncertainty in the occurrence of the event. Reliability, an overall measure of calibration, calculates the discrepancy between the observed outcome and its predictive probability. Resolution indicates the precision of the prediction. These three measures were corrected for sample size as suggested by Ferro *et al*. (227). For the sake of interpretation, these three components are usually combined into a single value representing the Brier score.

Calibration measures refer to the agreement between observed outcomes and predictions (105, 114). They can be assessed graphically using 1) a calibration plot, which is a simple scatter plot with predicted and observed values along the two axes, or 2) the Hosmer-Lemeshow goodness-of-fit test. However, the scatter plot is not appropriate for binary data, and the Hosmer-Lemeshow test has several drawbacks, including limited power and poor interpretability, as it only returns a chi-square value and *p*-value (105, 228).

Discrimination measures refer to the ability of a model to distinguish between subjects with and without the outcome of interest. These measures can be obtained by a contingency table, often called a confusion matrix or classification matrix, which can be used to classify problems where the outcome is binary or categorical. For a binary classification problem (such as disease presence or absence), a confusion matrix has two dimensions—one dimension is defined by the observed outcome of a subject, and the other is indexed by predicted outcome for the same subject estimated by a predictive model (**Table 5.1**).

Table 5.1: Confusion matrix for binary classification

	Predicted class: disease	Predicted class: no disease
Observed class: disease	True Positive (TP)	False Negative (FN)
Observed class: no disease	False Positive (FP)	True Negative (TN)

This matrix contains four possible situations, which enables us to derive the discrimination measures; the four counts are true positive (TP), true negative (TN), false positive (FP) and false negative (FN). True positives are the number of subjects with positive clinical outcomes correctly predicted. True negatives (TN) specify the number of subjects with negative clinical outcomes correctly predicted. False positives (FP) refer to the number of subjects with predicted positive clinical outcomes who were misclassified, as they actually had negative outcomes. False negatives (FN) refer to the number of subjects with predicted negative clinical outcomes who were misclassified. **Table 5.2** presents the most common measures for a binary classification based on the values of the confusion matrix.

Table 5.2: Discrimination measures obtained by confusion matrix

Measure	**Formula**	**Evaluation focus**
Sensitivity	$TP/(TP + FN)$	Effectiveness of a predictive model to identify subjects with the disease
Specificity	$TN/(TN + FP)$	Effectiveness of a predictive model to identify subjects without the disease
Positive predictive value	$TP/(TP + FP)$	Precision of predictive model to identify subjects with the disease
Negative predictive value	$TN/(TN + FN)$	Agreement between observed values with the negative values given by the predictive model

The most popular discriminative measures are sensitivity, specificity, predictive values and the area under the receiver operating characteristic (ROC) curve (105, 114, 115).

Sensitivity (true positive rate) describes the proportion of subjects with the disease who were correctly classified within the total group of subjects with the disease, whereas specificity (true negative rate) defines the proportion of subjects without the disease who were correctly identified within the total group of subjects without disease. In other words, sensitivity relates to the potential of a predictive model to recognise subjects with the disease, and specificity relates to its ability to recognise subjects without the disease. Sensitivity and specificity are considered to be independent of the disease prevalence, meaning that results are generalisable and that different studies with different disease prevalence can be compared. Positive predictive value (PPV) defines the probability that a subject who was diagnosed with the condition actually has the condition. Negative predictive value (NPV) describes the probability of not having the condition in a subject identified without the disease. Predictive values are influenced by disease prevalence; therefore, predictive values from one study should not be transferred to another setting with a different prevalence of the disease in the population.

The ROC curve graphically displays the trade-off between sensitivity and specificity and is useful for assigning the optimal cutoffs for clinical use. This cutoff represents the threshold for classifying subjects as having or not having the disease being studied that maximises the performance of the predictive model in terms of both sensitivity and specificity. A common method used to identify the optimal cut-off point is the Youden Index, which optimises the predictive model's discriminative ability, giving equal weight to sensitivity and specificity (229-231). In practice, this index defines a threshold for classifying subjects as diseased or non-diseased by maximising the potential performance of the predictive model in terms of both sensitivity and specificity. Therefore, subjects with an individual probability lower than the Youden Index are classified as healthy, while those with values above the threshold are classified as diseased.

Overall discriminative ability is often expressed as the area under the ROC curve (AUC), also known as the *c*-statistic, and indicates the probability that a randomly selected individual with the condition has a higher risk score than a randomly chosen subject without the condition. It has a value between 0.5 and 1; an AUC of 1 represents perfect prediction, and an AUC of 0.5 represents useless prediction (no better than flipping a coin) (105, 114, 115, 232). The AUC represents the ability of a model to

limit misclassifications, such as false positives and false negatives. The higher the AUC is, the better is the discriminative ability of the predictive model.

Likelihood ratios are measures of diagnostic accuracy and indicate the probability that a subject has or does not have the condition being analysed, given the results of a predictive model (115, 233). A likelihood ratio greater than one indicates that the test result is more likely to occur in subjects with the disease, whilst a likelihood ratio less than one indicates that the test result is more likely to occur in subjects without the disease (115, 233). Likelihood ratios above 10 and below 0.1 provide evidence to rule in or rule out the outcome (here, asthma), respectively (233).

5.2.3 Missing values in candidate predictors

Missing data in candidate predictors are common in epidemiological studies, and failing to account for them can lead to serious issues. Missing data can introduce bias in the parameter estimations, leading to poor reproducibly of the results (234, 235). Omitting subjects with missing values reduces sample size, which decreases statistical power (236) as well as predictive performance and increases standard error (234, 235). Missing values can occur for different reasons; for example, they were not collected by design; the respondent refused to answer and answers dropped out or respondents answered ‘don't know’.

Multiple imputation (MI) is a commonly used technique to infer the values of missing data (237); this procedure is usually preferable to a complete case analysis in which subjects with missing values are excluded from the analysis (104, 238, 239). The MI procedure assumes that the data are from a multivariate Normal distribution and that missing values are missing completely at random (MCAR) or missing at random (MAR) (237). In the MCAR assumption, missing values do not depend on observed or unobserved data, whereas in the MAR assumption, also referred to as missing conditionally at random, missing values do depend on observed information and can be explained by the other variables in the model. Although violation of the MAR assumption can be expected in most cases (240), research has shown that its violation does not seriously distort parameter estimates (1).

The process of MI starts by checking the normality assumptions and whether there is a pattern to the missing data (for example, subjects skip follow-up visits that are scheduled on Fridays). It then proceeds with creating a number (*m*) of completed datasets where missing values are imputed by using well-defined imputation models. The generation of *m* unique estimates for each missing value accounts for the uncertainty associated with imputed values. Each completed dataset is then analysed separately, resulting in *m* estimates for each parameter. Finally, results are combined using Rubin's Rule, which averages the *m* estimates from the imputed data to obtain a single estimate of the parameter. Then its corresponding standard error is calculated, incorporating within- and between-imputation uncertainty, that is, variability of parameter estimates for each imputed data point and discrepancy of estimates across imputed datasets. Current guidance recommends the use from five to 10 imputation datasets (*m*), according to the proportion of missing values, to obtain correct imputation models and reliable statistical inference (238). Next, I discuss in detail each step of the MI performed in the STELAR data.

I imputed missing values of candidate predictors within each STELAR cohort by using the Multivariate Imputation by Chained Equations (MICE) package in R (https://cran.r-project.org/web/packages/mice/mice.pdf). I examined the proportion of missing values as well as the missing data patterns among the 32 variables in each of the STELAR cohorts and then specified the imputation model for each variable that contained missing values. The imputation model for each variable was chosen based on the following guidelines: incomplete data in binary variables were imputed using a logistic regression model; predictive mean matching was used to impute missing values in continuous variables; and multinomial logistic regression was used for categorical variables with more than two levels. Predictive mean matching is widely used to produce imputations and is considered to perform better than procedures that assume normal distributions (241). This method imputes missing values by calculating a mean value from subjects who are most similar to the one with missing data, based on the distribution of the observed data. For each STELAR cohort, I created *m*=10 imputed datasets. The 10 imputed datasets of the Ashford and ASLPAC studies were then combined to create the development data, and the 10 imputed datasets of the other cohorts (IoW, MAAS and SEATON) were combined to produce the testing/validation data.

Statistical models were applied to each imputed dataset of the development data, and 10 sets of coefficients and standard errors for all predictors were obtained. Predictors were considered important and were included in the final predictive model when they were selected at least five times across the 10 imputed datasets (127). The regression coefficients were averaged across the 10 imputed datasets, and estimation of the standard errors was computed using Rubin's Rule (242).

Individual predictive probabilities were obtained for each imputed dataset of the validation data and then averaged across the 10 sets of imputed data to define an individual's risk of asthma and to assess the predictive ability of the models. Performance measures, such as AUC, were then obtained for each predictive model by using the Youden Index to identify the optimal cut-off point on the ROC curve and classify children as developing or not developing asthma in adolescence.

5.3 Results

5.3.1 Study population

In total, I developed the predictive models using information on 1,153 children between the ages of two and five years whose parents reported wheezing episode(s). I validated the models among 523 children with parent-reported wheezing symptoms between the ages of two and five years. These subjects at high risk of asthma at the age of 15–20 years include 1,008 from the ALSPAC, 145 from Ashford, 230 from IoW, 115 from MAAS and 178 from SEATON. The total numbers of subjects with asthma at age 15–20 years with available information on childhood wheezing symptoms were 139 (26%) and 90 (24%) in the development and validation data (**Table 5.3** and **Table 5.4**). Specifically, in the ALSPAC, Ashford, IoW, MAAS and SEATON studies, those numbers were 125 (26%), 14 (24%), 40 (19%), 27 (33%), 23 (25%) (**Table 5.3** and **Table 5.4**), respectively.

A large proportion of subjects reported wheezing symptoms, specifically, 193 (41.1%), 22 (36.7%), 78 (37.3%), 33 (41.3%), 33 (36.3%) in the ALSPAC, Ashford, IoW, MAAS and SEATON studies, respectively. Use of asthma medication was reported in fewer subjects (**Table 5.3** and **Table 5.4**). Thus, a number of subjects were classified as not having asthma because they answered positively to none or only one of the two following conditions for having asthma: wheezing in the past 12 months and asthma medication in the past 12 months.

It is important to note that at the time that most of these cohorts were seen before the age five years, many children who wheezed were more likely to receive asthma diagnosis and thus to start asthma medications. As a result, this may have determined a circularity between definition of asthma and study population leading to an increase in the prevalence of outcome.

Table 5.3: Definition of asthma at 15–20 years in children at high risk included in the development data

	ALSPAC	Ashford	Development data
	n (%)	n (%)	n (%)
Age 15–20 years			
Asthma	125 (26.3)	14 (23.7)	139 (26.0)
Current wheeze	193 (41.1)	22 (36.7)	215 (40.6)
Asthma treatment	165 (34.8)	15 (25.4)	180 (33.7)

Table 5.4: Definition of asthma at 15–20 years in children at high risk included in the validation data

	IoW	MAAS	SEATON	Validation data
	n (%)	n (%)	n (%)	n (%)
Age 15–20 years				
Asthma	40 (19.1%)	27 (33.3)	23 (25.0)	90 (23.6%)
Current wheeze	78 (37.3%)	33 (41.3)	33 (36.3)	144 (37.9%)
Asthma treatment	43 (20.6%)	30 (37.0)	34 (35.4)	107 (27.7%)

The proportion of missing values varied among candidate predictors and cohorts (**Table C.1**) with the highest proportions of missing values in information collected at recruitment and in the first years of life.

Among the STELAR cohorts, **Table 5.5** summarises the characteristics collected by age five years in children with wheeze who are at high risk of asthma in adolescence. Overall, children across cohorts had similar characteristics, although those in the development dataset were generally more likely to have eczema in early childhood, younger mothers and started weaning earlier than children in the validation dataset. Some questions were asked slightly differently among the STELAR cohorts, including those about frequency of wheezing and eczema. Therefore, I harmonised that information and derived new variables that were comparable among the STELAR cohorts (see Chapter 4 - Data sources, paragraph 4.3).

Table 5.5: Characteristics collected by age five years among the STELAR cohorts in high-risk population

Candidate predictor	ASLPAC n (%) or median (IQR)	Ashford n (%) or median (IQR)	IoW n (%) or median (IQR)	MAAS n (%) or median (IQR)	SEATON n (%) or median (IQR)
Sex (Female)	558 (55.4)	81 (55.9)	130 (56.5)	70 (60.9)	103 (57.9)
SES - professional/managerial occupations	260 (27.9)	27 (18.8)	20 (14.3)	49 (47.1)	67 (37.9)
Maternal asthma	185 (18.8)	33 (22.8)	30 (13.2)	40 (34.8)	46 (25.8)
Maternal eczema	308 (31.3)	29 (20.0)	35 (15.3)	25 (21.7)	49 (27.5)
Maternal hayfever	393 (39.9)	49 (33.8)	45 (19.6)	41 (35.7)	63 (35.4)
Paternal asthma	139 (15.7)	24 (16.6)	28 (12.2)	24 (20.9)	35 (19.7)
Paternal eczema	71 (8.1)	18 (12.4)	18 (7.8)	17 (14.8)	22 (12.4)
Older siblings	424 (75.6)	79 (54.5)	108 (57.1)	55 (47.8)	86 (48.3)
No. of older siblings	2 (1.0)	2 (2.0)	2 (1.0)	2 (1.0)	2 (1.0)
Birth order (first born)	137 (24.4)	66 (45.5)	84 (36.5)	55 (49.5)	92 (51.7)
Maternal age at recruitment	28 (7.0)	27 (6.3)	26 (7.8)	30 (7.5)	30 (5.7)
Premature birth	59 (8.8)	9 (6.2)	8 (3.5)	4 (3.5)	12 (7.1)
Ever breast fed	784 (78.1)	88 (60.7)	186 (81.2)	84 (73.0)	127 (71.3)
Vaginal delivery	883 (88.5)	119 (82.1)	68 (89.5)	81 (70.4)	127 (75.6)
Age at weaning (weeks)	13 (4.0)	13 (5.4)	14 (4.0)	15 (4.0)	14 (4.0)
Cat at 1 yr	396 (41.0)	58 (40.0)	72 (57.6)	20 (18.9)	42 (23.6)
Cat at 4/5 yrs	406 (43.1)	43 (29.7)	83 (36.1)	20 (17.5)	38 (21.3)
Dog at 1 yr	401 (41.5)	43 (29.7)	56 (44.8)	21 (19.8)	28 (15.7)
Dog at 4/5 yrs	411 (43.7)	33 (22.8)	60 (26.1)	17 (14.9)	29 (16.3)
Moved house by 2 yrs	187 (20.3)	22 (15.2)	24 (10.4)	30 (26.1)	22 (13.0)
Maternal smoking during pregnancy	185 (19.8)	39 (26.9)	60 (26.2)	13 (11.3)	26 (14.6)
Maternal smoking at 1 yr	503 (50.9)	43 (29.7)	93 (41.5)	23 (20.0)	21 (12.5)
Current wheeze at 1 yr	519 (51.5)	66 (45.5)	51 (22.2)	56 (48.7)	56 (31.5)
Current wheeze at 2/3 yrs	518 (48.1)	94 (64.8)	143 (62.2)	64 (55.7)	113 (63.5)
Current wheeze at 4/5 yrs	945 (93.8)	108 (74.5)	179 (77.8)	96 (83.5)	127 (71.3)
Frequency of wheezing at 1 yr					
Sometimes	319 (30.9)	57 (39.3)	41 (19.1)	5 (4.3)	38 (22.8)
Often	178 (17.2)	9 (6.2)	17 (7.9)	51 (44.3)	3 (1.8)
Frequency of wheezing at 2/3 yrs					
Sometimes	315 (32.2)	86 (59.3)	110 (47.8)	32 (27.8)	101 (56.7)
Often	192 (19.7)	8 (5.5)	33 (14.3)	31 (27.0)	8 (4.5)
Frequency of wheezing at 4/5 yrs					
Sometimes	579 (58.9)	97 (66.9)	89 (38.7)	26 (22.6)	104 (63.8)
Often	341 (34.7)	11 (7.6)	71 (30.9)	70 (60.9)	8 (4.9)
Eczema at 1 yr	411 (40.6)	60 (41.4)	21 (9.9)	49 (42.6)	51 (28.7)
Eczema at 2/3 yrs	404 (42.1)	65 (44.8)	60 (26.2)	45 (39.1)	73 (41.0)
Eczema at 4/5 yrs	403 (40.7)	74 (51.0)	65 (28.3)	64 (55.7)	105 (59.0)

Abbreviations - SES: socioeconomic status; yr(s): year(s).

Univariable (crude) associations of candidate predictors with asthma at age 15–20 years in the development dataset (ALSPAC and Ashford study) are reported in the supplementary tables (**Table C.2)**. In general, subjects with and without asthma in adolescence were similar with regard to most sociodemographic and environmental factors but differed significantly in terms of dog ownership, maternal history of allergic diseases and asthma-like symptoms in the first five years of life. Specifically, the prevalence of a doctor's diagnosis of eczema by age five years and of having atopic parents was higher in children with asthma. However, dog ownership at age one year was more likely to be reported in subjects without asthma than in those with asthma at age 15–20 years.

5.3.2 Model development

Results of the predictive models in the development dataset (ALSPAC and Ashford cohorts) derived from stepwise regression, LASSO, elastic net (EN), Bayesian and empirical Bayes methods after multiple imputation of missing values are reported in **Table 5.6**. For all the models, odds ratios (ORs) are reported only for the candidate predictors selected in one or more predictive models, and 95% confidence intervals (95% CI) and 95% credible intervals (95% CrI) are reported for the stepwise regression and Bayesian method, respectively.

In the final multivariable models developed (**Table 5.6**), the stepwise, LASSO, EN, Bayesian and empirical Bayes methods selected 11, 11, 11, 10 and 6 predictors out of 32. The classical regressions such as stepwise, LASSO and EN, selected the largest number of predictors, whilst the Bayesian method was the most parsimonious, with estimates of regression coefficients higher than the other methods.

Among all five predictive methods, eczema in early childhood, maternal history of allergic disease and the presence of a dog in the house were important predictors. Asthma between the ages of 15–20 years was also associated with being female, medium-low socioeconomic status, absence of a cat in the house and maternal smoking in early life.

Table 5.6: Analysis for asthma risk at 15–20 years in the high-risk population using a multiple imputation technique

Predictor	Age 15–20 yrs				
	Stepwise	**LASSO**	**EN**	**EB**	**Bayesian**
	OR[a] (95% CI[b])	**OR[a]**	**OR[a]**	**OR[a]**	**OR[a] (95% CrI[c])**
Sex (Female)	1.38 (0.92,2.09)	1.07	1.09	1.10	
SES (Partly-skilled occupations)	1.51 (0.95,2.39)	1.07	1.08	1.16	
Maternal asthma	1.77 (1.05,2.97)	1.31	1.31	1.39	1.61 (1.04,2.55)
Maternal eczema	0.67 (0.42,1.07)				
Maternal hayfever	1.53 (0.99,2.37)	1.25	1.25	1.29	1.57 (1.06,2.35)
Paternal asthma	1.66 (0.97,2.82)	1.09	1.09		
Cat at 1 yr	0.67 (0.44,1.03)	0.87	0.86	0.84	
Dog at 1 yr	0.51 (0.32,0.81)	0.71	0.72	0.67	0.60 (0.38,0.90)
Maternal smoking at 1 yr	1.45 (0.96,2.21)	1.06	1.08	1.12	
Eczema at 1 yr	1.62 (1.03,2.55)	1.23	1.24	1.27	1.66 (1.05,2.53)
Eczema at 2/3 yrs		1.25	1.26	1.25	1.69 (1.06,2.80)
Eczema at 4/5 yrs	2.10 (1.34,3.31)	1.60	1.56	1.65	2.13 (1.34,3.23)
No. of predictors selected	***11***	***11***	***11***	***10***	***6***

Abbreviations – Stepwise: stepwise regression; LASSO: least absolute shrinkage and selection operator; EN: elastic net; EB: empirical Bayes; SES: socioeconomic status; yr(s): year(s).

[a]Odds ratio (OR) of asthma at age 15–20 years for each candidate predictor, adjusting for all the others.

[b]95% confidence interval.

[c]95% credible interval.

5.3.3 External validation of model performance

In order to assess model performance, an individual's probability of developing asthma at age 15–20 years was predicted by the `regression coefficients of the models previously described, and the optimal cut-off of the ROC curve, derived using the Youden Index, was used to classify which individuals were predicted to develop asthma in adolescence. Performance measures of all the models from the main analysis, including sensitivity, specificity, predictive values and likelihood ratios, as well as their confidence intervals, are reported in **Table 5.7**.

Table 5.7: Performance measures for the analysis in the high-risk population using a multiple imputation technique

Performance measure	Age 15–20 years				
	Stepwise	**LASSO**	**EN**	**EB**	**Bayesian**
Sensitivity (95% CI)	0.79 (0.69,0.87)	0.64 (0.54,0.74)	0.64 (0.54,0.74)	0.64 (0.54,0.74)	0.60 (0.49,0.70)
Specificity (95% CI)	0.50 (0.44,0.56)	0.64 (0.58,0.69)	0.63 (0.57,0.68)	0.64 (0.58,0.69)	0.67 (0.61,0.73)
PPV (95% CI)	0.33 (0.26,0.39)	0.35 (0.28,0.43)	0.35 (0.28,0.43)	0.35 (0.28,0.43)	0.36 (0.28,0.44)
NPV (95% CI)	0.88 (0.83,0.93)	0.85 (0.80,0.90)	0.85 (0.80,0.90)	0.85 (0.80,0.90)	0.85 (0.79,0.89)
LR+ (95% CI)	1.57 (1.34,1.83)	1.78 (1.43,2.20)	1.73 (1.39,2.14)	1.78 (1.43,2.20)	1.83 (1.44,2.31)
LR– (95% CI)	0.43 (0.28,0.64)	0.56 (0.42,0.75)	0.57 (0.42,0.76)	0.56 (0.42,0.75)	0.60 (0.46,0.78)
Youden Index (95% CI)	0.29 (0.13,0.42)	0.28 (0.12,0.44)	0.27 (0.11,0.43)	0.28 (0.12,0.44)	0.27 (0.11,0.43)
AUC (95% CI)	0.66 (0.60,0.72)	0.67 (0.61,0.74)	0.67 (0.61,0.74)	0.68 (0.61,0.74)	0.67 (0.60,0.73)
Brier score	0.44	0.36	0.37	0.36	0.35
HL test (*p*-value)	>0.99	>0.99	>0.99	>0.99	0.84

Abbreviations – Stepwise: stepwise regression; LASSO: least absolute shrinkage and selection operator; EN: elastic net; EB: empirical Bayes; 95% CI: 95% confidence interval; PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR–: negative likelihood ratio; AUC: area under the receiver operating characteristic curve; HL: Hosmer-Lemeshow test.

Overall, the discriminative measures of the models evaluated in the validation data were comparable, with sensitivity ranging from 0.60 to 0.79, positive predictive value from 0.33 to 0.36, specificity from 0.50 to 0.67 and negative predictive value from 0.85 to 0.88. Although performance measures of LASSO, EN and EB were analogous, stepwise regression provided better sensitivity, negative predictive value and negative likelihood ratio than all other models. However, the Bayesian method, which resulted in the most parsimonious model, showed good predictive performance, with moderately higher specificity, positive predictive value and positive likelihood ratio compared to the other models and comparable discriminative ability as well as calibration.

As a trade-off between sensitivity and specificity, the AUC can be used to assess overall discriminative ability; this measure varied between 0.66 and 0.68, with stepwise regression being the least accurate model. The Brier scores varied from 0.36 to 0.44 suggesting moderate overall agreement between the predicted probabilities of later asthma and the observed values in the data. The Hosmer-Lemeshow showed *p*-value > 0.001, providing evidence that the models are a good fit (although this test has been highly criticized, as it tends to have low power and does not take overfitting into account).

5.4 Discussion

5.4.1 Main findings

The use of a large network of participants together with recently developed statistical approaches, such as the Bayesian regularization method, may have the potential to identify the most important asthma predictors and improve our ability to predict who might develop asthma. To my knowledge, this is the first study to predict asthma at age 15–20 years, making these results important for future research and studies. Evaluation of predictive performance by a wide range of measurements, such as discrimination, calibration and accuracy measures, and external validation of my predictive models demonstrated elevated generalisability of results in external populations showing potential for their applicability in clinical practice and medical research.

I applied several statistical methods to develop a predictive model in two prospective UK cohorts and externally validated them using information from three other birth cohort studies, in order to

predict asthma in adolescence among preschool children (age of two and five years) with wheezing symptoms. I included 32 candidate predictors collected by the age of five years, and my predictive models selected from 6 to 11 predictors representing variables such as sex, contact with animals, eczema in early childhood and parental history of allergies. The factors found to be important for predicting late asthma are in broad agreement with those of other prospective studies (141). Generally, advanced methods, such as EB and the Bayesian method, were more parsimonious than the stepwise and classical penalized regression methods, suggesting that more sophisticated techniques might be more suitable for identifying predictors associated with asthma development in adolescence. Results presented in this chapter show good external validity, with predictive ability (measured by the AUC) of over 65% among the five models, suggesting good generalisability to external populations. However, advanced statistical methods (EB and the Bayesian method) provided similar or moderate improvement in predictive performance compared to classical methods. Performance measures among LASSO, EN and EB were very similar. Stepwise regression provided better sensitivity, negative predictive value and negative likelihood ratio than all other models, while the Bayesian regularization method had higher specificity, positive predictive values and positive likelihood ratio compared to all other methods. Specifically, stepwise regression provided a sensitivity of 0.79 and positive predictive value of 0.33. Therefore, the chance that a subject is classified with late asthma is 79%, but among them only 33% truly have the condition. The Bayesian method provided a sensitivity of 0.60 and positive predictive value of 0.36, suggesting that this method has only 36% chance of identifying subjects who truly develop the disease at 15–20 years. Difference in predictive performance between stepwise and the other methods may be due to the type of predictors selected and their estimates. In fact, stepwise regression is the only method that excluded eczema at age two to three years but included maternal eczema in the final model.

Variation in predictive performance among methods may also depends by differences in proportion of false results (false negative and false positive), as they are included in the calculation of discriminative measures such as sensitivity and predictive values. The lower false negative are, the higher is sensitivity (**Table 5.2**). On the other hand, the lower false positive, the higher is specificity (**Table 5.2**). False-positive results occur when the predictive model forecasts a condition in subjects who ultimately do not develop it; false-negative results indicate individuals who are wrongly

predicted not to develop the condition that they do develop. The major costs of false-positive predictions include patients' or their caregivers' anxiety (which, itself, can lead to the development of other conditions) and the use of unnecessary treatments leading to avoidable side-effects and expenses. Conversely, the costs of false-negative predictions include delay in searching for medical help and inappropriate reassurance. These errors in identifying individuals can affect their health care costs and quality of life (243). Negative predictive value was high (over 85%) among all methods presented in this chapter, suggesting that they can correctly classify subjects without asthma and that the probability of false-negative results is low. Conversely, positive predictive value was low (below 36%), indicating a high chance of false-positive results and thus of misclassifying subjects who ultimately do not develop late asthma.

Given the data, no model stood out as being better at predicting asthma in adolescence. The seemingly obvious environmental as well as symptoms collected at five years might be not sufficient to figure out what will cause some wheezy children to develop asthma and others not to. Perhaps there are other factors such as genetic and immunological information that can describe the pathogenesis of such respiratory condition and predict this in adolescence.

5.4.2 Comparison with existing asthma predictive models

Several predictive models have been proposed by others to estimate the risk of late asthma among preschool children (85, 86, 90-92, 139, 140). Five of them aimed to validate existing predictive models in independent populations (90-92, 139, 140). The rest focused on developing new predictive models, and were internally validated (88, 106-108, 133) or not validated (87, 89, 94, 98, 131, 132, 134), leading to potentially poor generalisability and transportability of the results. The predictive models that have been externally validated were reviewed in Chapter 2 (Validation of childhood asthma predictive models: a systematic review). Of the methods that were externally validated, only one study [Pescatore *et al.* (106)] applied an advanced method, such as LASSO; this method was developed in children aged one to three years with parent-reported wheeze or chronic cough to predict asthma at seven to eight years. Pescatore's risk score tested 24 candidate predictors collected in the first three years of life, with only ten of them included in the final model.

The most important predictors were related to wheezing symptoms, such as shortness of breath, frequent wheeze, wheeze without colds, activity disturbance by wheeze, and wheeze/cough triggered by exercise. This risk score showed, at age eight years, high sensitivity, positive likelihood ratio and an AUC of 0.74. Similar results were obtained in the external validation study (140), suggesting high generabilizability of their findings. However, their risk score attempted to predict asthma at school age by using information collected at preschool age. Since their age at outcome assessment is different than the one in the present study, and predictive performance is highly dependent on age at outcome assessment, comparison between my model and those reported in Chapter 2 (Validation of childhood asthma predictive models: a systematic review) is not feasible. In fact, prediction is more challenging when the outcome is assessed farther into the future, compared to the collection time of predictors. Although asthma symptoms are common at preschool age, they frequently decline during school age but sometimes recur in adolescence (5, 6). Consequently, if physicians could predict whether children with asthma-like symptoms at preschool age would develop asthma in adolescence, such an early prognosis could prevent over-treatment of some children and help others take preventive steps to mitigate it, particularly if environmental factors are involved.

Chatzimichail and colleagues (106) achieved good predictive performance using retrospective data from 112 children who had a doctor's diagnosis of asthma in the Paediatric Department of the University Hospital of Alexandroupolis, Greece. In total, 46 predictors were included in the statistical model developed by using principal component analysis and least square support vector machine classifier methods. The evaluation of its performance was assessed by internal validation (10-fold cross-validation). In comparison with my predictive model, Chatzimichail *et al.* (106) included information that was not available in all STELAR cohorts and thus was not tested, such as number of family members, type of domestic heating, allergic conjunctivitis, dyspnoea, seasonal symptoms, breathing tests (FEV_1% and forced expiratory flow at 25–75% [FEF25/75]), IgE Units Per Millilitre [U/Ml] and pharmacological therapy (bronchodilators, corticosteroids, antileucotrines, antihistamines). That additional data, together with the use of advanced statistical methods, may have increased the sensitivity and specificity of their model compared to my predictive model.

However, their predictive model remains difficult to apply in clinical practice and to replicate in future studies, as some of the 46 predictors are difficult to collect. Furthermore, lack of external validation might have resulted in model overfitting and overly optimistic estimates of predictive performance. The lack of reporting measures, such as Brier score or AUC, makes it difficult to fully compare their predictive model with mine. Thus, the studies by Pescatore *et al*. and Chatzimichail *et al.*, along with results provided in this chapter, suggest that a suitable set of predictors and advanced statistical methods may be the next step forward for asthma prediction.

5.4.3 Strengths and limitations of this work

The main strength of this study is the prospective design of the five UK asthma birth cohorts and the length of their follow-up periods, from birth to adolescence. This allowed me to predict asthma at 15–20 years, which has not been done for those cohorts. To my knowledge, The STELAR data is the largest dataset of its kind, including over 20,000 children. Second, I tested a number of candidate predictors simultaneously by using advanced statistical methods; all of the information for those predictors can be obtained easily in clinical practice or by using a questionnaire. Third, I applied and compared different advanced regularization techniques within classical and Bayesian frameworks, to account for model overfitting and sample variability (125, 200, 244-248) and externally evaluated their predictive performance by using three 'plausible' populations within the STELAR consortium. An advanced method (the Bayesian model) was the most parsimonious method (six predictors) compared to the others (10-11 predictors) that also provided comparable predictive performance; hence, sophisticated techniques might be more appropriate to identify predictors associated to asthma at 15–20 years. Finally, I used the TRIPOD guideline (125) to report the main characteristics of my predictive models as well as their performance measures in order to allow for a fair comparison between my predictive models and existing models.

This study also had limitations. First, information was collected at comparable but not identical ages among the STELAR cohorts. The five UK birth cohorts were designed independent of each other and organized into a consortium post hoc; therefore, data on study children, their parents and their habitats were collected at different follow-up times.

Because children at similar ages have broadly similar characteristics and asthma-like symptoms, those differences likely have a low impact on the selection and estimation of the effects of predictors and thus on the performance of the predictive models (see **Table 5.5**). In addition, external validation enables one to generalize results to other populations, limiting potential differences among cohorts. Second, information about respiratory symptoms was collected through similarly worded questionnaires based on the International Study of Asthma and Allergies in Childhood (ISAAC) (165, 166); however, these self-reports reflect subjective interpretation of respiratory symptoms, which could lead to biased estimation of statistical results and poor transportability of findings. Nevertheless, the ISAAC questionnaire has been validated in different countries and populations, and it shows good repeatability and validity (249-253). Third, the long follow-up period resulted in fewer participating children. Fourth, asthma at 15-20 years was defined by both current wheezing and asthma medication leading, potentially, to the exclusion of individuals who do not wheeze due to their asthma treatment; this is likely to be a small proportion of individuals with a low impact on the performance of the predictive model. Finally, my study has was not validated outside the UK; therefore, my predictive models may have different predictive performance when tested in different countries.

5.5 Conclusion

There has been considerable interest in predictive models in modern clinical medicine, and policymakers are increasingly recommending their use in clinical practice guidelines (125). I presented several approaches that are able to estimate regression coefficients as well as perform variable selection for high-dimensional data. Logistic stepwise regression, LASSO, EN, Bayesian and empirical Bayes regularization methods were applied to predict asthma at age 15–20 years and to identify the best predictors for a large set of covariates. Variable selection and estimation of coefficients were described, and methods used for predictive performance (sensitivity, specificity, predictive values and AUC) were compared. Although the Bayesian method included just over half of the predictors included in the other models, that method had comparable predictive performance. Given that parsimonious models that have comparable performance measures are

usually preferred, the Bayesian method can be considered the best method for predicting the development of asthma between the ages of 15 and 20 years.

In the following chapter, I will explore whether prediction of asthma between age 15 and 20 years can be improved by including additional information collected in middle childhood. Specifically, I will apply only the Bayesian method to examine whether prediction ability can be improved by adding extra information collected from age five to 10 years, including respiratory symptoms and clinical measures, such as skin prick testing.

Chapter 6 – Further investigations using the Bayesian model

A number of asthma predictive models have been developed to forecast asthma in childhood, but they have shown limited clinical relevance. As discussed in Chapter 2, only a few of them have been externally validated and their predictive ability assessed in different populations with comparable characteristics in similar clinical settings. Those predictive models were developed and validated to forecast asthma in the first 10 years of life, and none of them showed outstanding predictive ability. None of the existing asthma predictive models predicted asthma in adolescence; by being able to predict the risk of a child developing asthma in adolescence, physicians could identify high-risk children and work to mitigate the development or progression of the disease. It would also be helpful to parents, who could attempt to reduce environmental exposures that exacerbate asthma. It would also assist in the design of, and recruitment to, studies of novel treatments. Conversely, accurately identifying children at low risk of developing asthma would help prevent the over-treatment of children with wheezing symptoms, thereby decreasing the incidence of side effects (83) and reducing costs of healthcare services (84).

Studies that aimed at identifying children who will likely develop asthma could be grouped into two main categories—studies of the general population recruited either at birth or through healthcare practices, schools or local registries and studies of children at high risk of developing late asthma, by recruiting infants born to allergic parents, approaching symptomatic children recruited at healthcare practices or schools or looking at children hospitalized for asthma symptoms. The API was the only model developed in a general population and externally validated in both the general population and in children at high risk of later asthma (see details in Chapter 2, Validation of childhood asthma

predictive models: a systematic review). Therefore, there is a need to develop a predictive model to predict asthma in adolescence in both populations.

Existing asthma predictive models tested a different set of candidate predictors collected at preschool age, making it difficult to compare models and to identify the most important predictors of later asthma. The existing asthma predictive models mainly assessed clinical characteristics collected at preschool age. Other potential predictors, such as environmental exposures and objective measurements, apart from blood eosinophilia (87), that are potentially associated with asthma development, were not considered in those models. The sole use of parent-reported information collected at preschool age, as for the analysis presented in the previous chapter of this PhD thesis, might be not sufficient to properly characterize either the health outcome or its candidate predictors. The need for objective measures, including more unbiased information, might be necessary to accurately disentangle predictions for children who have different disease entities that have respiratory symptoms similar to those of asthma and to explore the severity of such conditions.

This chapter will extend the findings of the previous chapter by using the Bayesian method to examine whether adding to the model extra information collected in mid-childhood, such as respiratory symptoms and some clinical measures (skin prick testing), could improve its predictive accuracy and clinical applicability to forecast asthma at age 15–20 years. These expanded models will be applied not only to symptomatic children considered to be at high risk of asthma later in life (specifically, those with wheezing symptoms at preschool age) but also to the full STELAR population, which includes all children recruited at birth among the cohorts.

6.1 Methods

6.1.1 Definition of target population and outcome

The analyses presented in this chapter were performed considering 1) all children recruited at birth among the five STELAR cohorts, to identify common characteristics within the general population; 2) a subgroup of study children with parent-reported wheeze between the ages of two and five years, who are at higher risk of asthma in adolescence.

The first group includes 14,451 children from the ALSPAC, 642 from Ashford, 1,456 from IoW, 1,184 from MAAS and 1,924 from SEATON, respectively, while the second group involves 1,008; 145; 230; 115 and 178 children from the same STELAR cohorts, respectively. Details on the STELAR cohorts can be found in Chapter 4 (Data sources, paragraph 4.1.1).

As reported in Chapter 5, current asthma was defined as the outcome and was assessed between 15 and 20 years of age, in order to develop models to predict the condition in adolescence. Full definition of asthma at 15–20 years can be found in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years, paragraph 5.1.2).

6.1.2 Potential predictive variables

In addition to the predictors described and used in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years, paragraph 5.1.3), a number of candidate predictors collected from age five to 10 years will be tested in this chapter, including wheezing symptoms between eight and 10 years, doctors' assessments and skin prick test results. Details on the harmonisation procedure as well as a list of candidate predictors can be found in Chapter 4 (Data sources, paragraph 4.3). As discussed in the previous chapter, missing values were imputed within the STELAR cohorts by using the MICE technique. Additional information on imputing missing values is described in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years, paragraph 5.2.3).

6.1.3 Statistical analysis

Similar to the statistical analysis described in the previous chapter, the present analysis was performed using a two-step approach: a development process followed by an evaluation of the predictive models (additional information is reported in Chapter 5 - Comparison of statistical methods for the prediction of asthma at 15–20 years, paragraph 5.2).

I developed four nested models that differ in terms of the candidate predictors considered, and I tested them in both general and high-risk populations:

- Model A included predictors collected by age five years (32 candidate predictors tested), and is the same model presented in Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years);
- Model B tested predictors collected by age five years and wheezing symptoms up to age 10 years (34 candidate predictors tested);
- Model C consisted of predictors tested in model B and asthma related information collected up to age eight, such as doctor's diagnosis of asthma and use of asthma medications (36 candidate predictors tested);
- Model D tested all predictors in model C and information collected by age 10 years including objective measurements such as skin prick test results (39 candidate predictors tested).

Finally, I evaluated the performance of those predictive models by measures of sensitivity, specificity, predictive values, area under the ROC curve and predictive values and calibration measures (additional information in Chapter 5, Comparison of statistical methods for the prediction of asthma at 15–20 years, paragraph 5.2.2).

As discussed in previous chapter, the ALSPAC and Ashford cohorts (development/training data) were used to develop the predictive models, and the other studies (validation/testing data) were used to externally validate their predictive abilities. **Figure 6.1** shows definition of the target populations and number of subjects among STELAR cohorts used for the statistical analysis.

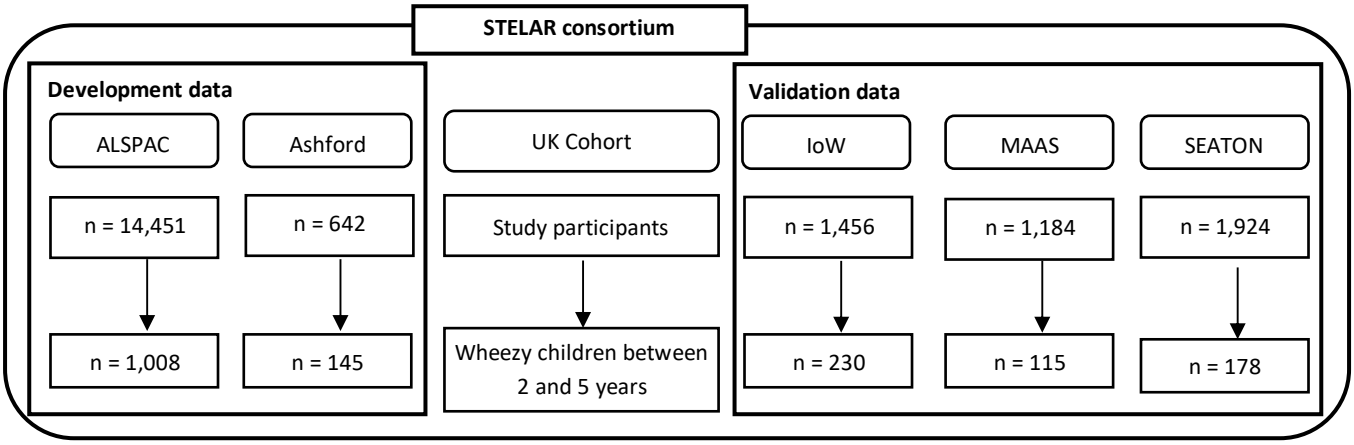


Figure 6.1: Definition of target populations in the STELAR consortium

6.2 Results

6.2.1 Study population

I developed the predictive models using information on 15,093 children from the general population and 1,153 children at high risk. In the general population (**Table 6.1**), 397 (7.9%) adolescents with complete information on the outcome had asthma at age 15–20 years, and in the high-risk population (**Table 5.3**) 139 (26.0%) did. The predictive models were validated in 4,564 and 523 children in the general population (**Table 6.2**) and subjects at high risk (**Table 5.4**), respectively. Among the general population, the number of adolescents with asthma at age 15–20 years was 302 (10.0%) (**Table 6.2**), while in the high-risk population it was 90 (23.6%) (**Table 5.4**).

Table 6.1: Definition of asthma at age 15–20 years in the general population included in the development data

Age 15–20 years	**ALSPAC**	**Ashford**	**Development data**
	n (%)	**n (%)**	**n (%)**
Asthma	375 (7.9)	22 (9.1)	397 (7.9)
Current wheeze	839 (17.6)	48 (19.6)	887 (17.7)
Asthma treatment	541 (11.3)	24 (10.0)	565 (11.2)

Table 6.2: Definition of asthma at age 15–20 years in the general population included in the validation data

Age 15–20 years	**IoW**	**MAAS**	**SEATON**	**Validation data**
	n (%)	**n (%)**	**n (%)**	**n (%)**
Asthma	143 (10.9%)	98 (13.0)	61 (8.6)	302 (10.9%)
Current wheeze	294 (6.6%)	127 (17.2)	115 (15.9)	536 (9.1%)
Asthma treatment	159 (12.1%)	128 (17.0)	96 (13.3)	383 (13.7%)

The proportions of missing values varied among candidate predictors and study cohorts (**Table D.1** and **Table D.2**). **Table 6.3** and **Table 6.4** summarise the characteristics of the general population and study children at high risk, respectively. Children across cohorts had similar characteristics; however, those in the development data and general population had less wheezing symptoms at 10 years and were more likely to have allergic mothers as well as younger mothers, to have moved house by age two years and started weaning earlier than children in the validation dataset. Children at high risk were more likely to be atopic, first born and to have at least one parent with an allergic disease (asthma, hayfever or eczema) compared to those in the general population.

Table 6.3: Characteristics collected by age 10 years among the STELAR cohorts in the general population

Candidate predictor	ASLPAC	Ashford	IoW	MAAS	SEATON
	n (%) or median (IQR)	n (%) or median (IQR)	n (%) or median (IQR)	n (%) or median (IQR)	n (%) or median (IQR)
Sex (Female)	7313 (51.7)	342 (53.4)	786 (51.2)	642 (54.2)	973 (50.5)
SES - professional/managerial occupations	3158 (26.4)	76 (12.0)	116 (16.0)	615 (61.2)	668 (34.0)
Maternal asthma	1437 (11.4)	90 (14.0)	145 (9.6)	231 (19.5)	334 (16.7)
Maternal eczema	2847 (22.6)	117 (18.2)	183 (12.1)	186 (15.7)	329 (16.4)
Maternal hayfever	3870 (30.7)	168 (26.2)	305 (20.1)	311 (26.3)	498 (24.8)
Paternal asthma	870 (9.0)	86 (13.4)	149 (9.9)	161 (13.6)	274 (13.7)
Paternal eczema	576 (6.0)	85 (13.3)	97 (6.4)	119 (10.1)	180 (9.0)
No. of older siblings	2 (1.0)	2 (1.0)	2 (1.0)	2 (1.0)	2 (1.0)
Birth order (first born)	1257 (23.5)	270 (42.1)	516 (42.4)	477 (44.3)	989 (49.3)
Maternal age at recruitment	28 (6.0)	28 (6.0)	26 (7.0)	31 (6.3)	30 (7.1)
Premature birth	858 (10.1)	44 (6.9)	49 (3.3)	43 (4.0)	151 (8.0)
Ever breast fed	9379 (73.7)	388 (62.3)	1058 (77.3)	788 (71.8)	1179 (70.7)
Vaginal delivery	6907 (82.3)	505 (79.0)	395 (91.0)	780 (76.5)	1448 (78.4)
Age at weaning (weeks)	13 (4.3)	13 (5.4)	14 (4.0)	15 (4.0)	15 (3.0)
Cat at 1 yr	4552 (41.7)	255 (39.7)	407 (56.6)	205 (20.0)	368 (22.3)
Cat at 4/5 yrs	4192 (44.2)	182 (22.4)	450 (28.9)	216 (15.4)	318 (17.6)
Dog at 1 yr	4678 (42.8)	183 (28.5)	353 (49.1)	174 (16.9)	289 (17.5)
Dog at 4/5 yrs	4131 (26.5)	144 (32.7)	352 (6.4)	168 (34.0)	298 (21.5)
Moved house by 2 yrs	1984 (19.8)	117 (18.8)	135 (11.1)	198 (16.7)	178 (11.7)
Maternal smoking during pregnancy	2645 (20.8)	135 (21.0)	393 (26.0)	119 (10.1)	352 (17.6)
Maternal smoking at 1 yr	6611 (49.5)	140 (22.5)	581 (43.1)	174 (14.7)	220 (14.5)
Current wheeze at 1 yr	4206 (34.7)	267 (42.9)	164 (12.2)	202 (39.1)	267 (16.2)
Current wheeze at 2/3 yrs	1336 (13.4)	189 (30.6)	183 (15.7)	242 (22.6)	213 (15.5)
Current wheeze at 4/5 yrs	1049 (12.1)	108 (17.9)	261 (21.4)	240 (22.5)	162 (12.9)
Current wheeze at 8 or 10 yrs	820 (9.9)	79 (13.3)	259 (18.9)	185 (18.0)	147 (15.8)
Frequency of wheezing at 1 yr					
Sometimes	2657 (22.0)	250 (40.1)	112 (8.8)	7 (1.4)	180 (11.1)
Often	834 (6.9)	17 (2.7)	51 (4.0)	195 (37.7)	12 (0.7)
Frequency of wheezing at 2/3 yrs					
Sometimes	1271 (12.5)	175 (28.4)	144 (12.4)	99 (9.2)	168 (12.2)
Often	439 (4.3)	14 (2.3)	39 (3.4)	137 (12.8)	17 (1.2)
Frequency of wheezing at 4/5 yrs					
Sometimes	636 (7.4)	97 (16.1)	109 (8.9)	88 (8.2)	123 (10.0)
Often	363 (4.2)	11 (1.8)	91 (7.5)	152 (14.2)	10 (0.8)
Frequency of wheezing at 8 or 10 yrs					
Sometimes	590 (7.2)	72 (12.1)	152 (11.6)	125 (12.2)	93 (10.1)
Often	187 (2.3)	7 (1.2)	107 (8.1)	59 (5.8)	48 (5.2)
Eczema at 1 yr	2935 (7.7)	204 (16.4)	80 (9.7)	176 (9.6)	356 (10.1)
Eczema at 2/3 yrs	2707 (26.5)	234 (37.9)	259 (21.1)	374 (34.9)	399 (29.0)
Eczema at 4/5 yrs	2258 (24.2)	230 (38.1)	279 (22.9)	504 (47.1)	524 (41.8)
DD asthma by 4/5 or 7.5 yrs	1656 (20.4)	118 (19.5)	182 (14.9)	252 (23.6)	146 (11.7)
Asthma treatment at 4/5 or 6.4 yrs	1226 (14.8)	98 (16.3)	177 (14.5)	254 (23.8)	145 (11.6)
SPT cat at 4/5 or 7 yrs	455 (6.8)	35 (6.4)	51 (5.2)	104 (10.9)	54 (7.7)
SPT house dust mite at 4/5 or 7 yrs	848 (12.6)	39 (7.1)	105 (10.7)	172 (18.0)	71 (10.1)
SPT mixed grass at 4/5 or 7 yrs	796 (11.8)	33 (6.0)	69 (7.1)	172 (18.0)	72 (10.3)

Abbreviations - SES: socioeconomic status; yr(s): years; DD: doctor's diagnosis; SPT: skin prick testing.

Table 6.4: Characteristics collected by age 10 years among the STELAR cohorts in high-risk population

Candidate predictor	ASLPAC n (%) or median (IQR)	Ashford n (%) or median (IQR)	IoW n (%) or median (IQR)	MAAS n (%) or median (IQR)	SEATON n (%) or median (IQR)
Sex (Female)	558 (55.4)	81 (55.9)	130 (56.5)	70 (60.9)	103 (57.9)
SES - professional/managerial occupations	260 (27.9)	27 (18.8)	20 (14.3)	49 (47.1)	67 (37.9)
Maternal asthma	185 (18.8)	33 (22.8)	30 (13.2)	40 (34.8)	46 (25.8)
Maternal eczema	308 (31.3)	29 (20.0)	35 (15.3)	25 (21.7)	49 (27.5)
Maternal hayfever	393 (39.9)	49 (33.8)	45 (19.6)	41 (35.7)	63 (35.4)
Paternal asthma	139 (15.7)	24 (16.6)	28 (12.2)	24 (20.9)	35 (19.7)
Paternal eczema	71 (8.1)	18 (12.4)	18 (7.8)	17 (14.8)	22 (12.4)
Older siblings	424 (75.6)	79 (54.5)	108 (57.1)	55 (47.8)	86 (48.3)
No. of older siblings	2 (1.0)	2 (2.0)	2 (1.0)	2 (1.0)	2 (1.0)
Birth order (first born)	137 (24.4)	66 (45.5)	84 (36.5)	55 (49.5)	92 (51.7)
Maternal age at recruitment	28 (7.0)	27 (6.3)	26 (7.8)	30 (7.5)	30 (5.7)
Premature birth	59 (8.8)	9 (6.2)	8 (3.5)	4 (3.5)	12 (7.1)
Ever breast fed	784 (78.1)	88 (60.7)	186 (81.2)	84 (73.0)	127 (71.3)
Vaginal delivery	883 (88.5)	119 (82.1)	68 (89.5)	81 (70.4)	127 (75.6)
Age at weaning (weeks)	13 (4.0)	13 (5.4)	14 (4.0)	15 (4.0)	14 (4.0)
Cat at 1 yr	396 (41.0)	58 (40.0)	72 (57.6)	20 (18.9)	42 (23.6)
Cat at 4/5 yrs	406 (43.1)	43 (29.7)	83 (36.1)	20 (17.5)	38 (21.3)
Dog at 1 yr	401 (41.5)	43 (29.7)	56 (44.8)	21 (19.8)	28 (15.7)
Dog at 4/5 yrs	411 (43.7)	33 (22.8)	60 (26.1)	17 (14.9)	29 (16.3)
Moved house by 2 yrs	187 (20.3)	22 (15.2)	24 (10.4)	30 (26.1)	22 (13.0)
Maternal smoking during pregnancy	185 (19.8)	39 (26.9)	60 (26.2)	13 (11.3)	26 (14.6)
Maternal smoking at 1 yr	503 (50.9)	43 (29.7)	93 (41.5)	23 (20.0)	21 (12.5)
Current wheeze at 1 yr	519 (51.5)	66 (45.5)	51 (22.2)	56 (48.7)	56 (31.5)
Current wheeze at 2/3 yrs	518 (48.1)	94 (64.8)	143 (62.2)	64 (55.7)	113 (63.5)
Current wheeze at 4/5 yrs	945 (93.8)	108 (74.5)	179 (77.8)	96 (83.5)	127 (71.3)
Current wheeze at 8 or 10 yrs	881 (9.9)	79 (13.3)	249 (18.3)	185 (18.0)	147 (15.8)
Frequency of wheezing at 1 yr					
Sometimes	319 (30.9)	57 (39.3)	41 (19.1)	5 (4.3)	38 (22.8)
Often	178 (17.2)	9 (6.2)	17 (7.9)	51 (44.3)	3 (1.8)
Frequency of wheezing at 2/3 yrs					
Sometimes	315 (32.2)	86 (59.3)	110 (47.8)	32 (27.8)	101 (56.7)
Often	192 (19.7)	8 (5.5)	33 (14.3)	31 (27.0)	8 (4.5)
Frequency of wheezing at 4/5 yrs					
Sometimes	579 (58.9)	97 (66.9)	89 (38.7)	26 (22.6)	104 (63.8)
Often	341 (34.7)	11 (7.6)	71 (30.9)	70 (60.9)	8 (4.9)
Frequency of wheezing at 8 or 10 yrs					
Sometimes	626 (7.1)	72 (12.4)	147 (11.8)	126 (12.2)	93 (10.1)
Often	205 (2.3)	7 (1.2)	102 (8.1)	60 (5.8)	47 (5.2)
Eczema at 1 yr	411 (40.6)	60 (41.4)	21 (9.9)	49 (42.6)	51 (28.7)
Eczema at 2/3 yrs	404 (42.1)	65 (44.8)	60 (26.2)	45 (39.1)	73 (41.0)
Eczema at 4/5 yrs	403 (40.7)	74 (51.0)	65 (28.3)	64 (55.7)	105 (59.0)
DD of asthma at 4/5 or 7.5 yrs	552 (61.0)	75 (51.7)	125 (54.3)	55 (47.8)	82 (47.4)
Asthma treatment at 4/5 or 6.4 yrs	594 (67.4)	74 (51.0)	117 (50.9)	72 (62.6)	87 (49.4)
SPT cat at 4/5 or 7 yrs	149 (22.5)	20 (15.4)	17 (8.0)	20 (19.6)	18 (15.5)
SPT house dust mite at 4/5 or 7 yrs	216 (32.3)	22 (16.8)	46 (21.7)	31 (30.4)	19 (16.4)
SPT mixed grass at 4/5 or 7 yrs	179 (26.8)	17 (13.1)	22 (10.4)	31 (30.4)	18 (15.5)

Abbreviations - SES: socioeconomic status; yr(s): years; DD: doctor's diagnosis; SPT: skin prick testing.

Univariable associations of candidate predictors with asthma at age 15–20 years in the development dataset among the general and high-risk populations are reported in the supplementary tables (**Table D.3** and **Table D.4)**. Overall, subjects with and without late asthma had similar sociodemographic and environmental characteristics in both target populations but differed significantly in terms of parental history of allergic diseases and asthma-like symptoms in the first 10 years of life. Specifically, in study subjects of both populations who had asthma in adolescence, the prevalence of sensitisation, atopic parents and a doctor's diagnosis of atopic disease (asthma and eczema) by age five or seven years, were much higher.

6.2.2 Predictive models

In the development dataset (ALSPAC and Ashford cohorts), results of the predictive models derived from the Bayesian method after multiple imputation of missing values are reported in **Table 6.5** and **Table 6.6**. For all four models, odds ratios (ORs) and their 95% credible intervals (95% CrI) are reported for the candidate predictors selected in one or more predictive models.

Overall, a smaller number of predictors was associated with asthma at age 15–20 years in the high-risk population compared to the general population. The final multivariable models developed in the general population (**Table 6.5**), selected 13, 13, 11 and 15 predictors, compared to 6, 5, 8 and 9 predictors for the high-risk population (**Table 6.6**). Models C and B were the most parsimonious models in both target populations followed by model D; in the general population, proportions of predictors selected among Model C, B and D are 30.5%, 38.2% and 38.5%, while in the high-risk population the proportions among those models are 22.2%, 14.7% and 23.1%.

Model A, which tested predictors collected by age five years, and Model B, which tested predictors by age five and wheezing symptoms between 8 and 10 years, are similar in terms of predictors included in the final model and their estimates, suggesting that wheezing symptoms collected between ages eight and 10 years do not improve the ability to predict asthma at 15–20 years. In addition, wheezing symptoms collected between eight and 10 years were not selected as predictors in most models in which that information was included, confirming the poor discriminative ability of wheezing at age eight to 10 years.

As expected, Models C and D, which included information on doctor's diagnosis of asthma, use of asthma medication and sensitization, showed that these are important predictors for asthma in adolescence; they represented the top predictors in both Model C and D in both target populations, suggesting a strong relationship with the outcome. For this reason, collecting that information in clinical practice may be useful to accurately distinguish patients who will from those who will not develop asthma in adolescence.

In summary, in the general population predictors selected in all models include sex, eczema in early childhood and parental allergies, whereas in the population of children at high risk important predictors among all models were parental allergy, eczema and dog in the house, with pet in the house being a protective factor for asthma at age 15–20 years. In both populations, allergen sensitization and doctor's diagnosis of asthma were the top predictors for the models that included them, suggesting that these factors have great potential to predict asthma in adolescence and to be used in clinical practice in order to evaluate the development of this condition. These top predictors can indeed be considered more objective and accurate than parent-reported information and can be easily assessed during clinical assessments, especially doctor's diagnosis of asthma.

Table 6.5: Analysis for asthma risk at age 15–20 years in the general population using multiple imputation technique

Predictor	Age 15–20 years			
	Model A	**Model B**	**Model C**	**Model D**
	OR[a] (95% CrI[b])	**OR[a] (95% CrI[b])**	**OR[a] (95% CrI[b])**	**OR[a] (95% CrI[b])**
Sex (Female)	1.33 (1.06,1.63)	1.39 (1.14,1.70)	1.51 (1.21,1.88)	1.80 (1.46,2.24)
SES (Skilled manual occupations)		0.51 (0.30,1.11)		
Maternal asthma	1.39 (1.05,1.89)	1.39 (1.06,1.74)		
Maternal hayfever	1.55 (1.25,1.95)	1.61 (1.30,2.00)	1.66 (1.33,2.10)	1.50 (1.21,1.88)
Paternal asthma	1.91 (1.44,2.51)	1.95 (1.51,2.56)	1.59 (1.22,2.13)	1.56 (1.18,2.09)
Older siblings				1.24 (0.98,1.69)
Maternal age at recruitment			0.95 (0.94,0.96)	0.95 (0.94,0.96)
Ever breast fed			1.53 (1.18,2.08)	1.66 (1.27,2.26)
Current wheeze at 2/3 yrs	1.59 (1.21,2.15)	1.73 (1.31,2.27)		
Current wheeze at 4/5 yrs	2.84 (2.07,4.15)	2.47 (1.72,3.52)		
Frequency of wheezing at 1 yr (Sometimes)	1.32 (1.03,1.63)			1.32 (1.06,1.63)
Frequency of wheezing at 2/3 yrs (Sometimes)	1.48 (1.13,1.91)	1.47 (1.10,1.96)		1.35 (1.04,1.75)
Frequency of wheezing at 2/3 yrs (Often)	1.66 (1.06,2.48)	1.60 (1.03,2.45)		
Frequency of wheezing at 4/5 yrs (Often)	1.94 (1.35,2.88)	2.91 (2.00,4.67)	1.65 (1.17,2.27)	
Eczema at 1 yr	1.59 (1.24,2.06)	1.42 (1.12,1.86)	1.40 (1.10,1.80)	1.36 (1.07,1.74)
Eczema at 2/3 yrs	1.44 (1.10,1.86)	1.62 (1.25,2.11)	1.59 (1.22,2.05)	1.37 (1.10,1.72)
Eczema at 4/5 yrs	1.59 (1.24,2.00)	1.43 (1.11,1.78)		
Current wheeze at 8 or 10 yrs	*Not included*		1.47 (1.00,2.05)	
DD of asthma at 4/5 or 7.5 yrs	*Not included*	*Not included*	3.57 (2.49,5.17)	3.00 (2.08,4.49)
Asthma treatment at 4/5 or 6.4 yrs	*Not included*	*Not included*	2.86 (2.03,4.02)	2.47 (1.75,3.70)
SPT cat at 4/5 or 7 yrs	*Not included*	*Not included*	*Not included*	2.05 (1.52,2.81)
SPT house dust mite at 4/5 or 7 yrs	*Not included*	*Not included*	*Not included*	2.02 (1.58,2.73)
SPT mixed grass at 4/5 or 7 yrs	*Not included*	*Not included*	*Not included*	2.11 (1.58,2.82)
No. of predictors selected	***13***	***13***	***11***	***15***

Abbreviations – SES: socioeconomic status; yr(s): year(s); DD: doctor's diagnosis; SPT: skin prick testing.
[a]Odds ratio (OR) of asthma at age 15–20 years for each candidate predictor, adjusting for all the others.
[b]95% credible interval.

Table 6.6: Analysis for asthma risk at 15–20 years in the high-risk population using a multiple imputation technique

Predictor	Age 15–20 yrs			
	Model A	**Model B**	**Model C**	**Model D**
	OR[a] (95% CrI[b])	**OR[a] (95% CrI[b])**	**OR[a] (95% CrI[b])**	**OR[a] (95% CrI[b])**
Sex (Female)			1.67 (1.14,2.46)	1.88 (1.25,2.90)
Maternal asthma	1.61 (1.04,2.55)	1.73 (1.12,2.78)	1.88 (1.09,2.89)	1.81 (1.09,3.26)
Maternal hayfever	1.57 (1.06,2.35)	1.64 (1.12,2.78)	1.92 (1.25,2.93)	1.86 (1.24,2.82)
Cat at 1 yr			0.60 (0.39,0.89)	0.58 (0.39,0.89)
Dog at 1 yr	0.60 (0.38,0.90)	0.61 (0.39,0.91)	0.52 (0.34,0.81)	0.54 (0.34,0.84)
Eczema at 1 yr	1.66 (1.05,2.53)			
Eczema at 2/3 yrs	1.69 (1.06,2.80)	1.81 (1.16,2.76)	2.05 (1.35,3.11)	1.82 (1.16,2.79)
Eczema at 4/5 yrs	2.13 (1.34,3.23)	1.73 (1.10,2.70)		
Current wheeze at 8 or 10 yrs	*Not included*	*Not included*	1.81 (1.22,2.84)	
DD of asthma at 4/5 or 7.5 yrs	*Not included*	*Not included*	5.54 (3.29,9.71)	4.84 (2.81,8.39)
SPT cat at 4/5 or 7 yrs	*Not included*	*Not included*	*Not included*	1.98 (1.19,3.24)
SPT house dust mite at 4/5 or 7 yrs	*Not included*	*Not included*	*Not included*	2.73 (1.60,4.48)
No. of predictors selected	***6***	***5***	***8***	***9***

Abbreviations – SES: socioeconomic status; yr(s): year(s); DD: doctor's diagnosis; SPT: skin prick testing.
[a]Odds ratio (OR) of asthma at age 15–20 years for each candidate predictor, adjusting for all the others.
[b]95% credible interval.

Performance measures of all the models from the main analysis, including sensitivity, specificity, predictive values and likelihood ratios, as well as their confidence intervals, are reported in **Table 6.7** and **Table 6.8**.

Table 6.7: Performance measures for the analysis in general population among models developed using a multiple imputation technique

Performance measure	Age 15–20 years			
	Model A	**Model B**	**Model C**	**Model D**
Sensitivity (95% CI)	0.68 (0.63,0.73)	0.56 (0.50,0.61)	0.66 (0.60,0.71)	0.56 (0.50,0.61)
Specificity (95% CI)	0.68 (0.66,0.70)	0.80 (0.78,0.81)	0.77 (0.75,0.78)	0.83 (0.81,0.84)
PPV (95% CI)	0.21 (0.18,0.23)	0.25 (0.22,0.29)	0.25 (0.22,0.29)	0.28 (0.25,0.32)
NPV (95% CI)	0.95 (0.93,0.96)	0.94 (0.93,0.95)	0.95 (0.94,0.96)	0.94 (0.93,0.95)
LR+ (95% CI)	2.11 (1.92,2.33)	2.76 (2.43,3.13)	2.79 (2.50,3.10)	3.22 (2.81,3.68)
LR– (95% CI)	0.47 (0.40,0.56)	0.56 (0.49,0.63)	0.45 (0.39,0.53)	0.54 (0.47,0.61)
Youden Index (95% CI)	0.36 (0.29,0.43)	0.35 (0.28,0.43)	0.42 (0.35,0.49)	0.38 (0.31,0.46)
AUC (95% CI)	0.73 (0.70,0.76)	0.73 (0.70,0.76)	0.76 (0.73,0.79)	0.75 (0.72,0.78)
Brier score	0.32	0.23	0.25	0.20
HL test (*p*-value)	>0.99	>0.99	>0.99	>0.99

Abbreviations – 95% CI: 95% confidence interval; PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR–: negative likelihood ratio; AUC: area under the receiver operating characteristic curve; HL: Hosmer-Lemeshow test.

Table 6.8: Performance measures for the analysis in the high-risk population among models developed using a multiple imputation technique

Performance measure	Age 15–20			
	Model A	**Model B**	**Model C**	**Model D**
Sensitivity (95% CI)	0.60 (0.49,0.70)	0.52 (0.41,0.63)	0.63 (0.53,0.73)	0.72 (0.61,0.81)
Specificity (95% CI)	0.67 (0.61,0.73)	0.72 (0.66,0.77)	0.78 (0.73,0.83)	0.70 (0.65,0.76)
PPV (95% CI)	0.36 (0.28,0.44)	0.36 (0.28,0.45)	0.47 (0.38,0.56)	0.43 (0.35,0.51)
NPV (95% CI)	0.85 (0.79,0.89)	0.83 (0.78,0.87)	0.87 (0.83,0.91)	0.89 (0.84,0.93)
LR+ (95% CI)	1.83 (1.44,2.31)	1.86 (1.42,2.44)	2.89 (2.21,3.78)	2.41 (1.93,3.01)
LR– (95% CI)	0.60 (0.46,0.78)	0.66 (0.53,0.83)	0.47 (0.36,0.62)	0.40 (0.29,0.57)
Youden Index (95% CI)	0.27 (0.11,0.43)	0.24 (0.08,0.40)	0.41 (0.25,0.56)	0.42 (0.26,0.56)
AUC (95% CI)	0.67 (0.60,0.73)	0.65 (0.58,0.71)	0.75 (0.70,0.81)	0.75 (0.70,0.81)
Brier score	0.35	0.33	0.26	0.29
HL test (*p*-value)	0.84	0.47	>0.99	>0.99

Abbreviations – 95% CI: 95% confidence interval; PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR–: negative likelihood ratio; AUC: area under the receiver operating characteristic curve; HL: Hosmer-Lemeshow test.

Overall, in both target populations, the discriminative measures of models in the validation data varied among models, with sensitivity ranging from 0.56 to 0.68 in the general population and from 0.52 to 0.72 in study children at high risk; the positive predictive value ranged from 0.21 to 0.28 and from 0.36 to 0.47 among the two target populations, respectively.

Generally, the models in the general population had better sensitivity and positive likelihood ratio, but positive predictive values were higher in models developed for subjects at high risk. These predictive measures such as sensitivity and positive predictive values, which are greatly needed in clinical practice to correctly identify subjects who are likely to develop the condition in adolescence, cannot yet be considered sufficient to accurately forecast asthma at 15–20 years. Although positive predictive value is twice as high in children at high risk compared to the general population, this measure still far from a value that can be considered truthful. Similarly, sensitivity is below 0.80 in both populations, suggesting that the model has poor ability to identify patients with the condition. A high value for these measures would allow physicians to accurately identify patients who will develop the condition in adolescence and would increase doctor-parent communication and help parents understand a long-term treatment management plan.

Specificity was similar among both populations but slightly higher in some models developed in the general population compared with those developed for high-risk children. Likelihood ratios were between 0.1 and 10 providing no evidence to rule in or rule out the subjects with asthma in adolescence. Performance measurements computed in the general population seemed to be more accurate (they had narrower confidence intervals), reflecting the larger sample size.

The Brier scores were higher in the models developed for children at high risk; this measure varied from 0.20 to 0.32 in the general population and from 0.26 to 0.35 in the high-risk population, suggesting good overall agreement between the predicted probabilities of later asthma and the observed values in the data. A lower Brier score indicates better calibration between the prediction and the actual outcome. This score was higher in Model A in both target populations; thus, this model was the least calibrated. The Hosmer-Lemeshow test provided a p-value > 0.001 suggesting that there is not enough evidence to say that the models are not a good fit.

Models that excluded skin prick testing information, doctor's diagnosis of asthma and/or use of asthma medication (Models A and B) had poorer predictive performance in both target populations compared to the other models. Models C and D showed similar predictive performance in both populations. Although they had higher sensitivity, positive likelihood ratio, predictive value and AUC compared to the other models, they are still far from perfect models. Their predictive ability and accuracy are limited, suggesting that additional information, such as genetic factors and clinical tests including lung function tests, allergen-specific IgE, fractional exhaled nitric oxide (FeNO) and bronchodilator response, might also be necessary to increase their ability to correctly identify children who will have asthma in adolescence.

6.3 Discussion

6.3.1 Asthma predictors

I applied the best variable selection model identified in the previous chapter, the Bayesian model, to develop and validate four versions of that model and to determine whether these expanded models would better predict asthma in adolescence. I used the same two STELAR cohorts (ALSPAC and Ashford) to develop the models and then externally validated them using the remaining three cohorts to predict asthma in adolescence among preschool children with and without wheezing symptoms between the age of two and five years. I tested a different set of candidate predictors, including demographic characteristics, wheezing symptoms, doctor's diagnosis, parental history of allergies and sensitisation to the most common allergens, and assessed that prediction can be improved by adding information collected in mid-childhood to the models.

Eczema in early childhood and parental allergies were important predictors in all models and both study populations. Sensitisation to common allergens and doctor's diagnosis of asthma were the top predictors in the models that included them among both populations, suggesting that this information might be useful in clinical practice to accurately distinguish patients who will develop asthma in adolescence. This information is indeed more accurate and less objective than parent-reported information and can be conveniently obtained during a standard clinical assessment.

My results are in accord with previous studies (162) showing that allergen sensitisation, parental allergies and eczema are linked to asthma development (294) and that the pathogenesis of asthma may be determined not only by the combination of these risk factors but also by inherited factors. Several epidemiological studies identified positive sensitisation to allergens as a risk factor as well as a predictor for asthma and childhood wheezing (141); skin prick tests and measurement of allergen-specific IgE antibodies are common diagnostic tools to confirm sensitisation (254). Specifically, inhalation of cat, dust or grass allergens release pro-inflammatory mediators that constrict the bronchioles. Exposure to allergens lead to chest tightness and wheezing symptoms and may result in narrowing of the airway of subjects and triggering allergic or atopic asthma, the most common type of asthma. Studies have shown that the proportions of children with and without asthma that is skin prick test–positive vary between different populations (255); in the STELAR cohorts, children at preschool age with wheezing symptoms were more likely to be atopic.

Although positive skin prick test results are known to be associated with the development of asthma, doctor's diagnosis of asthma and use of asthma medication in mid-childhood had not been tested in previous asthma predictive models. Thus, I tested those predictors for asthma at age 15–20 years. However, asthma was defined using information on wheezing symptoms and the use of asthma medication between ages 15 and 20 years; thus, the outcome definition may be correlated with use of asthma medication in mid-childhood. Nevertheless, use of asthma medication in mid-childhood was not selected as a predictor by any model developed for subjects at high risk. Additionally, inclusion of doctor's diagnosis of asthma (ever) in the definition of asthma in adolescence could have provided a more specific outcome definition and it could have been used to improve the predictive performance of models that do not include information on doctor's diagnosis of asthma as predictor (i.e. Models A and B). However, this additional asthma definition was not used in the analyses presented in this thesis to allow a reliable comparison across all four predictive models (i.e. Models A-D). My results indicate that the doctor's ability to correctly identify asthma at school age can be used 1) to improve screening as well as inclusion/exclusion criteria of future studies to assess the natural history of asthma, and 2) to identify subjects who are likely to develop asthma in adolescence and who have more need to receive treatment.

Inconsistency in the predictors selected among models may be caused by correlation between candidate predictors. Since the Bayesian regularization method selects only the most relevant predictors among those correlated, inclusion of mid-childhood information may have impacted on the selection of other predictors.

6.3.2 Predictive performance variation

In my study, predictive performance varied among models and target populations, with good but not sufficient overall predictive performance. Variation in predictive performance among models was determined by selecting different predictors leading to difference in their estimates. For example, information on wheezing symptoms in mid-childhood did not improve the predictive performance of the models, suggesting that asthma probably develops before age 10 years. Models that included objective or unbiased information, such as skin prick testing, doctor's diagnosis of asthma and/or use of asthma medication had higher predictive performance and better accuracy in both study populations. In general this information which were collected in mid-childhood increased the predictive performance of the models showing the variability in the expression of asthma-like symptoms during childhood and the unmet need to include more unbiased and/or objective information to predict asthma in adolescence. It worth noticing that information collected in mid-childhood such as SPT and doctor assessment, were gathered at similar but not identical ages among the STELAR cohorts with ALSPAC study reporting it around the age seven years and the others at five years. This study, which is the biggest across the STELAR cohorts may have driven the analysis; strong association of mid-childhood information with asthma in adolescence may have also depended by time proximity between this information and age of outcome assessment. Overall, model performance was still limited and needs to improve in order for the model to be clinically useful and accurately distinguish subjects who will develop asthma later in life.

Variation in predictive performance among target populations may have been determined not only by the definition of target populations, which could result in differences in the prevalence of the health outcome, but also by differences between STELAR cohorts, such as heterogeneity among subject characteristics, age at asthma assessment and collection time of candidate predictors. Several studies have shown that the prevalence of the health outcome influences predictive values, namely, discriminative measures(143, 154, 155). In general, limitations presented in Chapter 5

(Comparison of statistical methods for the prediction of asthma at 15–20 years) are also applicable here. Specifically, the STELAR cohorts were designed independent of one another; therefore, data were collected at different follow-up points, and subjects' characteristics may differ. Asthma in adolescence and candidate predictors harmonized were evaluated at comparable but not identical ages among the five cohorts.

Furthermore, information was collected using similar but not identical questionnaires eliciting subjective interpretation of symptoms. These differences might have increased the heterogeneity of subject characteristics between cohorts; the more different the characteristics among subjects and/or populations are the highest is the predictive performance variability. Assessment of the predictive performance of a model in a new sample of subjects from different locations and clinical settings is necessary; however, the validation cohort should be homogeneous with that of the development cohort in order to reduce variation in performance and to establish the generalisability of the results.

Decisions about which target population should be studied will depend on both research and clinical aims, efficacy and the possible side effects of therapeutic strategies. For example, general population studies focus on identifying prognostic factors and on creating and promoting healthy behavior and activities including environmental settings, diet and smoking habit (256-258). Population studies consider the entire range of subjects, as well as environmental and clinical factors and their interactions, that are associated with the development of the condition of interest in order to improve the health status of the population. Conversely, studies analysing symptomatic individuals focus on screening and identifying those at high risk of developing the condition of interest, to limit their exposure and/or to provide suitable interventions and personalised treatments (256-258). The early identification of children at high risk of developing asthma in adolescence is important during decision-making processes in both clinical practice and medical research. In clinical practice, such early detection would facilitate doctor-patient communication based on more objective information and would help physicians guide patients or their carers to decide on further testing and personalised treatments (105, 259). In medical research, the risk stratification of individuals would enable more accurate screening and establish proper inclusion and/or exclusion criteria for future epidemiological or experimental studies (105, 259).

6.3.3 Conclusion and implications for future research

Inclusion of information collected in mid-childhood improved predictive performance of the models and thus may be necessary to describe the natural history of asthma and to predict asthma later in life; in particular more objective and/or unbiased measures are needed to accurately distinguish children with or without asthma in adolescence.

Validation of the present results in cohorts from different countries is necessary to evaluate the reliability and generalisability of the results; therefore, future research is needed to test asthma predictive models in ongoing clinical trials and epidemiological studies around the world. Moreover, application of my predictive models in clinical trials would enable their predictive performance to be evaluated among different subgroups of patients that could potentially be established based on an individual's probability of having asthma in adolescence. Improved predictive performance could be achieved by incorporating not only physiologic measurements, such as lung function, bronchial hyper-responsiveness and exhaled nitric oxide, but also genetic risk factors into asthma-predictive models. However, obtaining such information may decrease the simplicity of predictive models for use in clinical practice.

Chapter 7 – General discussion

In clinical practice, the use of predictive models is common; they usually consist of a combination of medical assessments, clinical history evaluation and observation of symptoms. These models aim to assess a patient's health status in order to forecast the development, progression or remission of a given condition. Many prediction models have been proposed to forecast asthma in children, but their predictive performance, clinical usefulness and reproducibility in external populations have been limited, making them difficult to use in clinical practice and/or research. My aim was to develop better asthma prediction models, by comparing the performance of classical and advanced statistical methods and by developing and validating them in separate datasets. This PhD project was possible due to collaborative and multicentre efforts by the STELAR consortium. The five studies forming the consortium allowed for increased sample size, which I used to develop predictive models based on extensive data collected over 20 years, from birth to adolescence. Those data enabled me to develop models to predict asthma at age 15–20 years, which has not been done before. Thus, this project was planned to break new ground from both a clinical point of view and for future research.

7.1 Model validation in external studies and clinical settings

It is imperative that predictive models are validated in external studies and different clinical settings such as children with and without wheezing symptoms at preschool age, as this enables their predictive ability to be quantified and their generalizability assessed. Several studies also highlight the need to distinguish between population-based and high-risk individual-based approaches for assessing population- and individual-level health problems (256, 257). Whereas the population-based approach aims to decrease the risk of the disease in the entire population by promoting healthy behaviours and supporting prevention policies and interventions, the individual-based

approach identifies high-risk individuals to offer them a suitable prevention or management strategy.

In the research conducted for this thesis, I completed a systematic review of existing asthma predictive models that have been externally validated. For the literature on asthma prediction, I explored their predictors, definitions of target outcome as well as target populations, evaluated their ability to predict asthma in later life and identified research gaps that needed to be addressed. Overall, the number of validated asthma predictive models is limited, and they provide poor predictive accuracy, with some performance variation in sensitivity and positive predictive value. In total, only three prediction models were externally validated and therefore considered in this review; only Pescatore *et al.* (108) developed models using advanced statistical methods such as LASSO regression.

This PhD project also evaluated model predictive performance in two target populations — a subgroup of symptomatic children with wheezing at preschool age (high-risk population), and the full STELAR population (general population) described in Chapter 6 (Further investigations using the Bayesian model). Those analyses allowed me to examine the consistency of the models that I developed and to explore the importance of assessing predictive ability across different situations that can occur in clinical practice. Predictive ability can be enhanced by studying general and high-risk populations in parallel in the same study, to assess the model predictive performance in different clinical scenarios. In Chapter 6, I showed that the number of selected predictors, as well as their performance measures, varied by population. In the general population, more predictors than in the high-risk population were selected, mainly wheezing symptoms and their frequency; however, these additional 'predictors' are used to define the high-risk population, so that the selected predictors were compatible in the two populations. It was impossible to compare the performance of my models with the existing predictive models assessed in Chapter 2 (Validation of childhood asthma-predictive models: a systematic review), as none of them considered the same target population, age of asthma assessment and type of predictors.

7.2 Assessment of statistical methods

In Chapter 5 (Comparison of statistical methods for the prediction of asthma at 15–20 years), I investigated the use of classical and modern techniques to perform model selection and predict asthma between ages 15 and 20 years by using information harmonized in early childhood (up to five years of age) among the STELAR cohorts. Overall, the advanced approaches, such as empirical Bayes and Bayesian methods, were more parsimonious than stepwise and classical penalized regression, suggesting that more sophisticated techniques might be more suitable for identifying predictors associated with asthma in adolescence.

However, classical and advanced methods performed similarly in the validation data, suggesting that inclusion of further information, such as genetic factors and lung function measurements, may be more crucial than the use of sophisticated statistical approaches to improve the performance of future asthma prediction models. In addition, advanced regularisation methods are known to be more powerful and robust than classical statistical approaches, but the difference is more marked when applied to high-dimensional data where classical methods suffer from overfitting. Therefore, similarity in predictive performance among my statistical methods may be due to the relatively low-dimensional data of the explored target population. In this chapter, the data used to validate the statistical methods cannot be considered 'high-dimensional' as the number of candidate predictors was relatively low.

7.3 Predictive performance in practice

In Chapter 6 (Further investigations using the Bayesian model), I used the Bayesian method to examine additional candidate predictors to assess whether information collected in mid-childhood (up to 10 years of age) could improve the predictive accuracy and clinical applicability of models developed to forecast asthma at age 15–20 years. Although mid-childhood information increased the predictive performance of the models especially when objective measures were included, their predictive ability was still limited.

My predictive models have only moderate accuracy and predictive performance. In daily practice, physicians and their patients are more interested in positive and negative predictive values, as these measures represent the probability of being correctly classified after a test. However, predictive values have only limited value in assessing the predictive ability of a model as they depend on the prevalence of the disease and therefore on the composition of the population being evaluated (114, 115). Sensitivity and specificity are usually the benchmarks to assess the predictive performance of models. Determining a threshold at which the predictive model has a 'satisfactory' accuracy and deciding whether high sensitivity is preferred over specificity or vice versa have important clinical impacts, but they are not straightforward (114). Generally, a screening test should be highly sensitive, while a confirmatory test should be highly specific (260). Screening tests are usually conducted to identify groups of subjects with similar risk profiles, whereas confirmatory tests are used to validate a patient's health status. The predictive measures of my models were assessed using the optimal cut-off on the ROC curve in order to maximise both sensitivity and specificity, and accuracy was measured by the AUC. This approach to the evaluation of predictive measures and discriminative ability has the advantage of not being affected by arbitrary and subjective decision criteria.

In theory, a useful predictive model would be characterized by high sensitivity and specificity with accuracy approaching 100%, thereby correctly identifying subjects who will have or not have the condition. In practice, a perfect predictive model is unlikely to be achieved, especially when the forecasted outcome is a long time into the future (as with my prediction of asthma 15 to 20 years hence). Long-term prediction of future health status is usually challenging, as it requires reliable early-life information and relatively stable or progressive disease throughout the predictive time frame (104, 118, 119). Because the body can indeed heal, the longer the range and the more heterogeneous the condition is over time, the more difficult it is to forecast the outcome.

7.4 Outcome and predictors characterization

The lack of an effective and precise predictive model to identify children who will have asthma in adolescence may be due to others factors, including misspecification of the outcome, asthma, the absence of relevant risk factors in the predictive model, and heterogeneity across cohorts.

To minimise misspecification of the outcome, the UK National Institute of Health and Care Excellence (NICE) proposed an asthma diagnostic algorithm which starts with a spirometry test to identify airway obstruction; the bronchodilator reversibility test is then considered to confirm the obstruction and its reversibility; finally the FeNO and peak flow variability tests are performed to confirm the spirometry findings. However, performing these examinations is generally difficult in preschool children, making the diagnosis problematic and unreliable until after age five years (6). Physicians are thus advised to assess the condition based on symptoms and clinical assessment when children are unable to perform these tests.

The definition of asthma mainly based on symptoms and self-reported information, might contribute to misdiagnosis of asthma. An incorrect diagnosis of asthma can occur by misinterpreting respiratory symptoms not related to asthma and/or self-reported information regarding the term 'wheeze' (9); for example, it is possible that people mistake being 'unable to catch their breath' as 'wheezing'. Misdiagnosis of a health outcome such as asthma can affect the reliability and performance of a predictive model, leading to bias in estimated regression coefficients. Consequently, the model will wrongly predict an individual's risk of asthma later in life and produce inaccurate predictive performance measures.

Many childhood conditions have respiratory symptoms that resemble asthma; they include respiratory sinus infections, enlarged lymph nodes, congenital immune deficiency, metapneumovirus, parainfluenza, coronavirus and other viral infections. In epidemiological studies, the collection of information on asthma-like symptoms, diagnosis of asthma and drug treatment generally relies on ISAAC's core questions, which are completed by children or their parents/caregivers. Although that questionnaire has been validated in different countries (167, 168, 250), studies have shown that parents are likely to over-report severe symptoms, with fathers describing fewer symptoms than mothers, and adolescents tending to over-report asthma symptoms, including doctors' diagnoses, compared to their parents (261-263). In addition, diagnosis and treatment of asthma may be influenced by the accessibility of healthcare services.

Symptoms may be difficult for non-clinical individuals to assess; for example, family members of children who have never wheezed might not be familiar with 'wheezing', using this term to describe instead a wide range of noises and sensations (8, 264).

Overall, the use of respiratory symptoms and response to asthma treatments reported by non-clinicians, as for the main analysis in this PhD thesis, might not be sufficient to properly characterize either the health outcome or the candidate predictors. The need for objective measures, including lung function tests, might be necessary to accurately distinguish children with similar respiratory symptoms but different disease entities and to explore the severity of such conditions. The implications of inadequate definitions of outcome and risk factors, based on parental- (or self-) reported information, include unreliability of data, misinterpretation of findings and, in the context of predictive models, poor generalisability and lack of clinical usefulness.

In Chapter 6 (Further investigations using the Bayesian model), several predictors, including wheezing symptoms, eczema in early childhood, history of parental allergies, allergic sensitisation to common allergens and a clinical diagnosis of asthma, were associated with asthma in adolescence; that information had been collected prospectively through questionnaires and skin prick tests and then harmonised across the five STELAR cohorts. Although my results are in line with previous studies, information obtained in mid-childhood, such as doctor's diagnosis of asthma and use of asthma medication, had not been assessed as candidate predictors in previous asthma predictive models. These variables were identified in my models as important predictors with high effect estimates; however, there may be a circularity in their use as 'predictors', collected in mid-childhood, to forecast asthma just 10-12 years later. Specifically, doctor's diagnosis and asthma medication collected at 4/5 or 6.4/7.5 years were used as potential predictors of asthma at 15–20 years, defined as concurrent wheezing and asthma treatment. Whilst doctor diagnosis was not incorporated in the definition of asthma at 15–20 years, it is likely that a doctor diagnosis at school age would be correlated with wheezing and asthma treatment in adolescence. On the other hand, asthma treatment at 4/5 or 7 years was not consistently selected by all models presented in this thesis.

The ALSPAC study, which is the biggest study within the STELAR consortium, reported this information at around age seven years, while all other cohorts reported it around age five years

showing heterogeneity among cohorts and potentially leading to imprecise and unreliable estimates. Specifically, since the collection time of such mid-childhood information in ALSPAC study was closer to age of outcome assessment than the one gathered in the other cohorts, the selection as well as estimates of such information might have been driven by the ALSPAC study. Nevertheless, its predictive effect was mitigated by the external validation process assed in the other three STELAR cohorts. The selection of this information may also highlight a doctor's ability to correctly identify children in early childhood who are more likely to develop asthma in adolescence and are in greater need for asthma treatment. This ability could be considered more objective than simple information collected though questionnaires and thus, might be important to explore in future research.

7.5 Additional considerations

As shown in this PhD project wheezing symptoms and early-onset allergic sensitisation were considered essential for the prediction of asthma in adolescence. These factors have been consistently linked to asthma development (265), and children with both of those symptoms have been identified as being at particularly high risk of having asthma at school age (7, 266). Common triggers for asthma include exposure to infection or aeroallergens, seasonal or perennial, and wheezing occurring in the first five years of life, usually predominant during the autumn and winter months. Such triggers are almost exclusively associated with lower respiratory tract infections determined by common viruses (7, 267, 268).

Several studies linked persistent wheezing episodes to allergen sensitisation and doctor's diagnosis of eczema (269, 270), indicating that the pathogenesis of asthma might in part be determined by a combination of those risk factors and by a deviation of the immune system toward a particular cytokine response. Numerous studies also examined the relationship between production of cytokines response in atopic dermatitis and showed that impairment between T helper cells might affect the development of the skin condition in early life (271-273). This insight may explain associations and common biological mechanisms between respiratory and skin disorders.

The interactions between genetic susceptibility, environmental stressors, immune response and respiratory symptoms in early life might shape the progression and outcome of airway disease; therefore, examination of such interactions could enable us to understand the mechanisms affecting

the development of asthma and to identify the critical periods during which the condition starts. I explored a variety of risk factors, but others, such as genetic factors, cytokine levels and microbiome exposures, were not available across all STELAR studies and thus could not be considered, leading to the potential exclusion of key information. Although inclusion of those factors in predictive models may enhance the forecasting of asthma in adolescence, they may not be readily available to doctors thus limiting the application in clinical practice of predictive models that include them.

References

1. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330-51.
2. A plea to abandon asthma as a disease concept. Lancet. 2006;368(9537):705.
3. Asthma: diagnosis, monitoring and chronic asthma management. NICE Guideline. 2018.
4. Sullivan P, Ghushchyan VG, Navaratnam P, Friedman HS, Kavati A, Ortiz B, et al. School absence and productivity outcomes associated with childhood asthma in the USA. J Asthma. 2018;55(2):161-8.
5. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and wheezing in the first six years of life. The Group Health Medical Associates. N Engl J Med. 1995;332(3):133-8.
6. Murray C, Foden P, Lowe L, Durrington H, Custovic A, Simpson A. Diagnosis of asthma in symptomatic children based on measures of lung function: an analysis of data from a population-based birth cohort study. Lancet Child Adolesc Health. 2017;1(2):114-23.
7. Saglani S. Viral infections and the development of asthma in children. Ther Adv Infect Dis. 2013;1(4):139-50.
8. Bush A, Fleming L. Diagnosis and management of asthma in children. BMJ. 2015;350:h996.
9. Lau S, Nickel R, Niggemann B, Gruber C, Sommerfeld C, Illi S, et al. The development of childhood asthma: lessons from the German Multicentre Allergy Study (MAS). Paediatr Respir Rev. 2002;3(3):265-72.
10. Bush A, Pavord ID. Towards the abolition of asthma? Asthma Commission podcast: Lancet. 2017.
11. Los H, Postmus PE, Boomsma DI. Asthma genetics and intermediate phenotypes: a review from twin studies. Twin Res. 2001;4(2):81-93.
12. Miller RL, Ho SM. Environmental epigenetics and asthma: current concepts and call for studies. Am J Respir Crit Care Med. 2008;177(6):567-73.
13. 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(13):1211-21.
14. Prescott SL. Early-life environmental determinants of allergic diseases and the wider pandemic of inflammatory noncommunicable diseases. J Allergy Clin Immunol. 2013;131(1):23-30.
15. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nat Med. 2012;18(5):716-25.
16. Deliu M, Belgrave D, Sperrin M, Buchan I, Custovic A. Asthma phenotypes in childhood. Expert Rev Clin Immunol. 2017;13(7):705-13.
17. Henderson AJ. Childhood asthma phenotypes in the twenty-first century. Breathe. 2014;10 (2).
18. Wenzel SE. Asthma: defining of the persistent adult phenotypes. Lancet. 2006;368(9537):804-13.
19. Martinez FD. Development of wheezing disorders and asthma in preschool children. Pediatrics. 2002;109(2 Suppl):362-7.
20. Silverman M, Kuehni CE. Early lung development and COPD. Lancet. 2007;370(9589):717-9.
21. Brand PL, Baraldi E, Bisgaard H, Boner AL, Castro-Rodriguez JA, Custovic A, et al. Definition, assessment and treatment of wheezing disorders in preschool children: an evidence-based approach. Eur Respir J. 2008;32(4):1096-110.
22. 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(5):974-81.
23. Just J, Saint-Pierre P, Gouvis-Echraghi R, Boutin B, Panayotopoulos V, Chebahi N, et al. Wheeze phenotypes in young children have different courses during the preschool period. Ann Allergy Asthma Immunol. 2013;111(4):256-61 e1.
24. 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(1):103-10 e8.

25. Schultz A, Brand PL. Phenotype-directed treatment of pre-school-aged children with recurrent wheeze. J Paediatr Child Health. 2012;48(2):E73-8.
26. Blair H. Natural history of childhood asthma. 20-year follow-up. Arch Dis Child. 1977;52(8):613-9.
27. Granell R, Henderson AJ, Sterne JA. Associations of wheezing phenotypes with late asthma outcomes in the Avon Longitudinal Study of Parents and Children: A population-based birth cohort. J Allergy Clin Immunol. 2016;138(4):1060-70 e11.
28. De Sario M, Di Domenicantonio R, Corbo G, Forastiere F, Pistelli R, Rusconi F, et al. Characteristics of early transient, persistent, and late onset wheezers at 9 to 11 years of age. J Asthma. 2006;43(8):633-8.
29. Belgrave DC, Custovic A, Simpson A. Characterizing wheeze phenotypes to identify endotypes of childhood asthma, and the implications for future management. Expert Rev Clin Immunol. 2013;9(10):921-36.
30. Henderson J, Granell R, Heron J, Sherriff A, Simpson A, Woodcock A, 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(11):974-80.
31. Belgrave DCM, Simpson A, Semic-Jusufagic A, Murray CS, Buchan I, Pickles A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. J Allergy Clin Immunol. 2013;132(3):575-83 e12.
32. Savenije OE, Granell R, Caudri D, Koppelman GH, Smit HA, Wijga A, et al. Comparison of childhood wheezing phenotypes in 2 birth cohorts: ALSPAC and PIAMA. J Allergy Clin Immunol. 2011;127(6):1505-12 e14.
33. Depner M, Fuchs O, Genuneit J, Karvonen AM, Hyvarinen A, Kaulek V, et al. Clinical and epidemiologic phenotypes of childhood asthma. Am J Respir Crit Care Med. 2014;189(2):129-38.
34. Stein RT, Martinez FD. Asthma phenotypes in childhood: lessons from an epidemiological approach. Paediatr Respir Rev. 2004;5(2):155-61.
35. Sears MR, Greene JM, Willan AR, Wiecek EM, Taylor DR, Flannery EM, et al. A longitudinal, population-based, cohort study of childhood asthma followed to adulthood. N Engl J Med. 2003;349(15):1414-22.
36. Phelan PD, Robertson CF, Olinsky A. The Melbourne Asthma Study: 1964-1999. J Allergy Clin Immunol. 2002;109(2):189-94.
37. Collins SA, Pike KC, Inskip HM, Godfrey KM, Roberts G, Holloway JW, et al. Validation of novel wheeze phenotypes using longitudinal airway function and atopic sensitization data in the first 6 years of life: evidence from the Southampton Women's survey. Pediatr Pulmonol. 2013;48(7):683-92.
38. 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(5):573-8.
39. Hallit S, Leynaert B, Delmas MC, Rocchi S, De Blic J, Marguet C, et al. Wheezing phenotypes and risk factors in early life: The ELFE cohort. PLoS One. 2018;13(4):e0196711.
40. Fallucca S, Vasta M, Sciullo E, Balducci S, Fallucca F. Birth weight: genetic and intrauterine environment in normal pregnancy. Diabetes Care. 2009;32(12):e149.
41. Hayatbakhsh MR, Sadasivam S, Mamun AA, Najman JM, Williams GM, O'Callaghan MJ. Maternal smoking during and after pregnancy and lung function in early adulthood: a prospective study. Thorax. 2009;64(9):810-4.
42. Turner S. Perinatal programming of childhood asthma: early fetal size, growth trajectory during infancy, and childhood asthma outcomes. Clin Dev Immunol. 2012;2012:962923.
43. Allan KM, Prabhu N, Craig LC, McNeill G, Kirby B, McLay J, et al. Maternal vitamin D and E intakes during pregnancy are associated with asthma in children. Eur Respir J. 2015;45(4):1027-36.
44. Douros K, Moustaki M, Tsabouri S, Papadopoulou A, Papadopoulos M, Priftis KN. Prenatal Maternal Stress and the Risk of Asthma in Children. Front Pediatr. 2017;5:202.
45. Castro-Rodriguez JA, Forno E, Rodriguez-Martinez CE, Celedon JC. Risk and Protective Factors for Childhood Asthma: What Is the Evidence? J Allergy Clin Immunol Pract. 2016;4(6):1111-22.
46. Bloomfield FH, Oliver MH, Harding JE. The late effects of fetal growth patterns. Arch Dis Child Fetal Neonatal Ed. 2006;91(4):F299-304.
47. Wignarajah D, Cock ML, Pinkerton KE, Harding R. Influence of intrauterine growth restriction on airway development in fetal and postnatal sheep. Pediatr Res. 2002;51(6):681-8.
48. Dell S, To T. Breastfeeding and asthma in young children: findings from a population-based study. Arch Pediatr Adolesc Med. 2001;155(11):1261-5.

49. Sears MR, Greene JM, Willan AR, Taylor DR, Flannery EM, Cowan JO, et al. Long-term relation between breastfeeding and development of atopy and asthma in children and young adults: a longitudinal study. Lancet. 2002;360(9337):901-7.
50. Wright AL, Holberg CJ, Taussig LM, Martinez F. Maternal asthma status alters relation of infant feeding to asthma in childhood. Adv Exp Med Biol. 2000;478:131-7.
51. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol. 2010;125(2 Suppl 2):S3-23.
52. Vock C, Hauber HP, Wegmann M. The other T helper cells in asthma pathogenesis. J Allergy (Cairo). 2010;2010:519298.
53. Romagnani S. T-cell subsets (Th1 versus Th2). Ann Allergy Asthma Immunol. 2000;85(1):9-18; quiz , 21.
54. Becklake MR, Kauffmann F. Gender differences in airway behaviour over the human life span. Thorax. 1999;54(12):1119-38.
55. Sears MR. Growing up with asthma. BMJ. 1994;309(6947):72-3.
56. Nicolai T, Illi S, Tenborg J, Kiess W, v Mutius E. Puberty and prognosis of asthma and bronchial hyper-reactivity. Pediatr Allergy Immunol. 2001;12(3):142-8.
57. Meurer JR, George V, Subichin S, Yauck J, Layde P. Asthma severity among children hospitalized in 1990 and 1995. Arch Pediatr Adolesc Med. 2000;154(2):143-9.
58. Skobeloff EM, Spivey WH, St Clair SS, Schoffstall JM. The influence of age and sex on asthma admissions. JAMA. 1992;268(24):3437-40.
59. Trawick DR, Holm C, Wirth J. Influence of gender on rates of hospitalization, hospital course, and hypercapnea in high-risk patients admitted for asthma : a 10-year retrospective study at Yale-New Haven Hospital. Chest. 2001;119(1):115-9.
60. Chen Y, Stewart P, Johansen H, McRae L, Taylor G. Sex difference in hospitalization due to asthma in relation to age. J Clin Epidemiol. 2003;56(2):180-7.
61. Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol. 2016;16(10):626-38.
62. Yung JA, Fuseini H, Newcomb DC. Hormones, sex, and asthma. Ann Allergy Asthma Immunol. 2018;120(5):488-94.
63. Schaub B, Lauener R, von Mutius E. The many faces of the hygiene hypothesis. J Allergy Clin Immunol. 2006;117(5):969-77; quiz 78.
64. Matricardi PM, Franzinelli F, Franco A, Caprio G, Murru F, Cioffi D, et al. Sibship size, birth order, and atopy in 11,371 Italian young men. J Allergy Clin Immunol. 1998;101(4 Pt 1):439-44.
65. Goldberg S, Israeli E, Schwartz S, Shochat T, Izbicki G, Toker-Maimon O, et al. Asthma prevalence, family size, and birth order. Chest. 2007;131(6):1747-52.
66. Bernsen RM, de Jongste JC, van der Wouden JC. Birth order and sibship size as independent risk factors for asthma, allergy, and eczema. Pediatr Allergy Immunol. 2003;14(6):464-9.
67. Halfon N, Newacheck PW. Childhood asthma and poverty: differential impacts and utilization of health services. Pediatrics. 1993;91(1):56-61.
68. Miller JE. The effects of race/ethnicity and income on early childhood asthma prevalence and health care use. Am J Public Health. 2000;90(3):428-30.
69. Kozyrskyj AL, Ernst P, Becker AB. Increased risk of childhood asthma from antibiotic use in early life. Chest. 2007;131(6):1753-9.
70. Kummeling I, Stelma FF, Dagnelie PC, Snijders BE, Penders J, Huber M, et al. Early life exposure to antibiotics and the subsequent development of eczema, wheeze, and allergic sensitization in the first 2 years of life: the KOALA Birth Cohort Study. Pediatrics. 2007;119(1):e225-31.
71. Devereux G, Barker RN, Seaton A. Antenatal determinants of neonatal immune responses to allergens. Clin Exp Allergy. 2002;32(1):43-50.
72. Cookson W, Moffatt M. Making sense of asthma genes. N Engl J Med. 2004;351(17):1794-6.
73. Ober C. Asthma Genetics in the Post-GWAS Era. Ann Am Thorac Soc. 2016;13 Suppl 1:S85-90.
74. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet. 2011;43(9):887-92.
75. Grotenboer NS, Ketelaar ME, Koppelman GH, Nawijn MC. Decoding asthma: translating genetic variation in IL33 and IL1RL1 into disease pathophysiology. J Allergy Clin Immunol. 2013;131(3):856-65.
76. Tsitsiou E, Williams AE, Moschos SA, Patel K, Rossios C, Jiang X, et al. Transcriptome analysis shows activation of circulating CD8+ T cells in patients with severe asthma. J Allergy Clin Immunol. 2012;129(1):95-103.

77. Martinez FD. Gene-environment interactions in asthma: with apologies to William of Ockham. Proc Am Thorac Soc. 2007;4(1):26-31.
78. Romieu I, Sienra-Monge JJ, Ramirez-Aguilar M, Moreno-Macias H, Reyes-Ruiz NI, Estela del Rio-Navarro B, et al. Genetic polymorphism of GSTM1 and antioxidant supplementation influence lung function in relation to ozone exposure in asthmatic children in Mexico City. Thorax. 2004;59(1):8-10.
79. Kim JK, Samaranayake M, Pradhan S. Epigenetic mechanisms in mammals. Cell Mol Life Sci. 2009;66(4):596-612.
80. Reik W, Dean W. DNA methylation and mammalian epigenetics. Electrophoresis. 2001;22(14):2838-43.
81. Kanno Y, Vahedi G, Hirahara K, Singleton K, O'Shea JJ. Transcriptional and epigenetic control of T helper cell specification: molecular mechanisms underlying commitment and plasticity. Annu Rev Immunol. 2012;30:707-31.
82. Hollingsworth JW, Maruoka S, Boon K, Garantziotis S, Li Z, Tomfohr J, et al. In utero supplementation with methyl donors enhances allergic airway disease in mice. J Clin Invest. 2008;118(10):3462-9.
83. Luskin AT, Antonova EN, Broder MS, Chang EY, Omachi TA, Ledford DK. Health care resource use and costs associated with possible side effects of high oral corticosteroid use in asthma: a claims-based analysis. Clinicoecon Outcomes Res. 2016;8:641-8.
84. Nunes C, Pereira AM, Morais-Almeida M. Asthma costs and social impact. Asthma Res Pract. 2017;3:1.
85. Luo G, Nkoy FL, Stone BL, Schmick D, Johnson MD. A systematic review of predictive models for asthma development in children. BMC Med Inform Decis Mak. 2015;15:99.
86. Smit HA, Pinart M, Anto JM, Keil T, Bousquet J, Carlsen KH, et al. Childhood asthma prediction models: a systematic review. Lancet Respir Med. 2015;3(12):973-84.
87. Castro-Rodriguez JA, Holberg CJ, Wright AL, Martinez FD. A clinical index to define risk of asthma in young children with recurrent wheezing. Am J Respir Crit Care Med. 2000;162(4 Pt 1):1403-6.
88. Caudri D, Wijga A, CM AS, Hoekstra M, Postma DS, Koppelman GH, et al. Predicting the long-term prognosis of children with symptoms suggestive of asthma at preschool age. J Allergy Clin Immunol. 2009;124(5):903-10.e1-7.
89. Chang TS, Lemanske RF, Guilbert TW, Gern JE, Coen MH, Evans MD, et al. Evaluation of the modified asthma predictive index in high-risk preschool children. J Allergy Clin Immunol Pract. 2013;1(2):152-6.
90. Devulapalli CS, Carlsen KC, Haland G, Munthe-Kaas MC, Pettersen M, Mowinckel P, et al. Severity of obstructive airways disease by age 2 years predicts asthma at 10 years of age. Thorax. 2008;63(1):8-13.
91. Hafkamp-de Groen E, Lingsma HF, Caudri D, Levie D, Wijga A, Koppelman GH, et al. Predicting asthma in preschool children with asthma-like symptoms: validating and updating the PIAMA risk score. J Allergy Clin Immunol. 2013;132(6):1303-10.
92. Rodriguez-Martinez CE, Sossa-Briceno MP, Castro-Rodriguez JA. Discriminative properties of two predictive indices for asthma diagnosis in a sample of preschoolers with recurrent wheezing. Pediatr Pulmonol. 2011;46(12):1175-81.
93. Guilbert TW, Morgan WJ, Krawiec M, Lemanske RF, Jr., Sorkness C, Szefler SJ, et al. The Prevention of Early Asthma in Kids study: design, rationale and methods for the Childhood Asthma Research and Education network. Control Clin Trials. 2004;25(3):286-310.
94. Amin P, Levin L, Epstein T, Ryan P, LeMasters G, Khurana Hershey G, et al. Optimum predictors of childhood asthma: persistent wheeze or the Asthma Predictive Index? J Allergy Clin Immunol Pract. 2014;2(6):709-15.
95. 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(3):163-9.
96. Matricardi PM, Illi S, Gruber C, Keil T, Nickel R, Wahn U, et al. Wheezing in childhood: incidence, longitudinal patterns and factors predicting persistence. Eur Respir J. 2008;32(3):585-92.
97. Vial Dupuy A, Amat F, Pereira B, Labbe A, Just J. A simple tool to identify infants at high risk of mild to severe childhood asthma: the persistent asthma predictive score. J Asthma. 2011;48(10):1015-21.
98. van der Mark LB, van Wonderen KE, Mohrs J, van Aalderen WMC, ter Riet G, Bindels PJE. Predicting asthma in preschool children at high risk presenting in primary care: Development of a clinical asthma prediction score. Prim Care Respir J. 2014;23(1):52-9.
99. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
100. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56(5):441-7.

101. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595.
102. Lee YH, Bang H, Kim DJ. How to Establish Clinical Prediction Models. Endocrinol Metab (Seoul). 2016;31(1):38-44.
103. Grobman WA, Stamilio DM. Methods of clinical prediction. Am J Obstet Gynecol. 2006;194(3):888-94.
104. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
105. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-31.
106. Chatzimichail E, Paraskakis E, Sitzimi M, Rigas A. An intelligent system approach for asthma prediction in symptomatic preschool children. Computational and mathematical methods in medicine. 2013;2013:240182.
107. van der Werff SD, Junco Diaz R, Reyneveld R, Heymans MW, Ponce Campos M, Gorbea Bonet M, et al. Prediction of asthma by common risk factors: a follow-up study in Cuban schoolchildren. J Investig Allergol Clin Immunol. 2013;23(6):415-20.
108. Pescatore AM, Dogaru CM, Duembgen L, Silverman M, Gaillard EA, Spycher BD, et al. A simple asthma prediction tool for preschool children with wheeze or cough. Journal of Allergy & Clinical Immunology. 2014;133(1):111-8.e1-13.
109. Refaeilzadeh P., Tang L. LH. Cross-Validation. 2009;Encyclopedia of Database Systems. Springer.
110. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-81.
111. Lisboa PJ, Taktak AFG. The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks. 2006;19(4):408-15.
112. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European Heart Journal. 2017;38(23):1805-14.
113. Harrell FE. Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis. Springer. 2001.
114. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models A Framework for Traditional and Novel Measures. Epidemiology. 2010;21(1):128-38.
115. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012;13 Suppl 4:S2.
116. Custovic A, Ainsworth J, Arshad H, Bishop C, Buchan I, Cullinan P, et al. The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together. Thorax. 2015;70(8):799-801.
117. Masoli M, Fabian D, Holt S, Beasley R, Global Initiative for Asthma P. The global burden of asthma: executive summary of the GINA Dissemination Committee report. Allergy. 2004;59(5):469-78.
118. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer Science & Business Media. 2008 Dec 16.
119. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
120. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
121. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10(2):e1001380.
122. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
123. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009;338:b604.
124. Wyatt JC, Altman DG. Prognostic Models - Clinically Useful or Quickly Forgotten - Commentary. Brit Med J. 1995;311(7019):1539-41.
125. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
126. Hui J, Oka A, James A, Palmer LJ, Musk AW, Beilby J, et al. A genome-wide association scan for asthma in a general Australian population. Hum Genet. 2008;123(3):297-306.

127. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27(17):3227-46.
128. Zhou DX. On grouping effect of elastic net. Statistics & Probability Letters. 2013;83(9):2108-12.
129. Zhang R, Borisenko O, Telegina I, Hargreaves J, Ahmed AR, Sanchez Santos R, et al. Systematic review of risk prediction models for diabetes after bariatric surgery. Br J Surg. 2016;103(11):1420-7.
130. Ivanescu AE, Li P, George B, Brown AW, Keith SW, Raju D, et al. The importance of prediction model validation and assessment in obesity and nutrition research. Int J Obes (Lond). 2016;40(6):887-94.
131. Cano-Garcinuno A, Mora-Gandarillas I, Group SS. Wheezing phenotypes in young children: an historical cohort study. Prim. 2014;23(1):60-6.
132. Ro AD, Simpson MR, Storro O, Johnsen R, Videm V, Oien T. The predictive value of allergen skin prick tests and IgE tests at pre-school age: The PACT study. Pediatric Allergy and Immunology. 2014;25(7):691-8.
133. Eysink PED, ter Riet G, Aalberse RC, van Aalderen WMC, Roos CM, van der Zee JS, et al. Accuracy of specific IgE in the prediction of asthma: Development of a scoring formula for general practice. British Journal of General Practice. 2005;55(511):125-31.
134. Boersma NA, Meijneke RWH, Kelder JC, van der Ent CK, Balemans WAF. Sensitization predicts asthma development among wheezing toddlers in secondary healthcare. Pediatric Pulmonology. 2017;52(6):729-36.
135. National Asthma E, Prevention P. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. J Allergy Clin Immunol. 2007;120(5 Suppl):S94-138.
136. Colicino S, Munblit D, Minelli C, Custovic A, Cullinan P. Validation of childhood asthma predictive tools: A systematic review. Clin Exp Allergy. 2019;49(4):410-8.
137. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744.
138. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6(7):e1000100.
139. Leonardi NA, Spycher BD, Strippoli MPF, Frey U, Silverman M, Kuehni CE. Validation of the Asthma Predictive Index and comparison with simpler clinical prediction rules. Journal of Allergy and Clinical Immunology. 2011;127(6):1466-72.e6.
140. Grabenhenrich LB, Reich A, Fischer F, Zepp F, Forster J, Schuster A, et al. The novel 10-item asthma prediction,tool: external validation in the german,MAS Birth cohort. PLoS ONE. 2014;9(12).
141. Rodriguez-Martinez CE, Sossa-Briceno MP, Castro-Rodriguez JA. Factors predicting persistence of early wheezing through childhood and adolescence: a systematic review of the literature. J Asthma Allergy. 2017;10:83-98.
142. Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. Can Med Assoc J. 2013;185(11):E537-E44.
143. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol. 2009;62(1):5-12.
144. Looijmans-van den Akker I, van Luijn K, Verheij T. Overdiagnosis of asthma in children in primary care: a retrospective analysis. Br J Gen Pract. 2016;66(644):e152-7.
145. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
146. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG, Cochrane IPDM-aMg. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 2015;12(10):e1001886.
147. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32(18):3158-80.
148. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137(7):598-602.
149. Willis BH. Spectrum bias--why clinicians need to be cautious when applying diagnostic test studies. Fam Pract. 2008;25(5):390-6.
150. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ. 2016;353:i3139.

151. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract. 2006;12(2):132-9.
152. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994;271(9):703-7.
153. Guyatt G, Sackett DL, Haynes RB. Evaluating diagnostic tests. Clinical epidemiology: how to do clinical practice research Lippincott williams & wilkins. 2012 Mar 29.
154. Feinstein AR. Misguided efforts and future challenges for research on "diagnostic tests". J Epidemiol Community Health. 2002;56(5):330-2.
155. Brenner HE, Gefeller OL. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Statistics in medicine. 1997 May 15;16(9):981-91.
156. Marks GB. Identifying asthma in population studies: from single entity to a multi-component approach. Eur Respir J. 2005;26(1):3-5.
157. Belgrave D, Simpson A, Custovic A. Challenges in interpreting wheeze phenotypes: the clinical implications of statistical learning techniques. Am J Respir Crit Care Med. 2014;189(2):121-3.
158. Belsky DW, Sears MR, Hancox RJ, Harrington H, Houts R, Moffitt TE, et al. Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. Lancet Respir Med. 2013;1(6):453-61.
159. Atkinson W, Harris J, Mills P, Moffat S, White C, Lynch O, et al. Domestic aeroallergen exposures among infants in an English town. European Respiratory Journal. 1999;13(3):583-9.
160. Atkinson W, Harris J, Mills P, Moffat S, White C, Lynch O, et al. Domestic aeroallergen exposures among infants in an English town. Eur Respir J. 1999;13(3):583-9.
161. Martindale S, McNeill G, Devereux G, Campbell D, Russell G, Seaton A. Antioxidant intake in pregnancy in relation to wheeze and eczema in the first two years of life. Am J Respir Crit Care Med. 2005;171(2):121-8.
162. Custovic A, Simpson BM, Simpson A, Hallam C, Craven M, Brutsche M, et al. Manchester Asthma and Allergy Study: low-allergen environment can be achieved and maintained during pregnancy and in early life. J Allergy Clin Immunol. 2000;105(2 Pt 1):252-8.
163. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2013;42(1):111-27.
164. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013;42(1):97-110.
165. Asher MI, Keil U, Anderson HR, Beasley R, Crane J, Martinez F, et al. International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. Eur Respir J. 1995;8(3):483-91.
166. Asher MI, Weiland SK. The International Study of Asthma and Allergies in Childhood (ISAAC). ISAAC Steering Committee. Clin Exp Allergy. 1998;28 Suppl 5:52-66; discussion 90-1.
167. Sole D, Vanna AT, Yamada E, Rizzo MC, Naspitz CK. International Study of Asthma and Allergies in Childhood (ISAAC) written questionnaire: validation of the asthma component among Brazilian children. J Investig Allergol Clin Immunol. 1998;8(6):376-82.
168. Valle SO, Kuschnir FC, Sole D, Silva MA, Silva RI, Da Cunha AJ. Validity and reproducibility of the asthma core International Study of Asthma and Allergies in Childhood (ISAAC) written questionnaire obtained by telephone survey. J Asthma. 2012;49(4):390-4.
169. Groves RM. Survey errors and survey costs. John Wiley & Sons. 2004.
170. Bradburn NM. Presidential address: a response to the nonresponse problem. 1992.
171. van Gelder MM, Bretveld RW, Roeleveld N. Web-based questionnaires: the future in epidemiology? Am J Epidemiol. 2010;172(11):1292-8.
172. Greenlaw C, Brown-Welty S. A comparison of web-based and paper-based survey methods: testing assumptions of survey mode and response cost. Eval Rev. 2009;33(5):464-80.
173. Morton LM, Cahill J, Hartge P. Reporting participation in epidemiologic studies: a survey of practice. Am J Epidemiol. 2006;163(3):197-203.
174. Etter JF, Perneger TV. Analysis of non-response bias in a mailed health survey. J Clin Epidemiol. 1997;50(10):1123-8.
175. Zuidgeest M, Hendriks M, Koopman L, Spreeuwenberg P, Rademakers J. A comparison of a postal survey and mixed-mode survey using a questionnaire on patients' experiences with breast care. J Med Internet Res. 2011;13(3):e68.

176. McMahon SR, Iwamoto M, Massoudi MS, Yusuf HR, Stevenson JM, David F, et al. Comparison of e-mail, fax, and postal surveys of pediatricians. Pediatrics. 2003;111(4 Pt 1):e299-303.
177. Marcano Belisario JS, Jamsek J, Huckvale K, O'Donoghue J, Morrison CP, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. Cochrane Database Syst Rev. 2015(7):MR000042.
178. Scott A, Jeon SH, Joyce CM, Humphreys JS, Kalb G, Witt J, et al. A randomised trial and economic evaluation of the effect of response mode on response rate, response bias, and item non-response in a survey of doctors. BMC Med Res Methodol. 2011;11:126.
179. Kongsved SM, Basnov M, Holm-Christensen K, Hjollund NH. Response rate and completeness of questionnaires: a randomized study of Internet versus paper-and-pencil versions. J Med Internet Res. 2007;9(3):e25.
180. Touvier M, Mejean C, Kesse-Guyot E, Pollet C, Malon A, Castetbon K, et al. Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. Eur J Epidemiol. 2010;25(5):287-96.
181. van den Berg MH, Overbeek A, van der Pal HJ, Versluys AB, Bresters D, van Leeuwen FE, et al. Using web-based and paper-based questionnaires for collecting data on fertility issues among female childhood cancer survivors: differences in response characteristics. J Med Internet Res. 2011;13(3):e76.
182. Wyatt JC. When to use web-based surveys. J Am Med Inform Assoc. 2000;7(4):426-9.
183. Tenforde AS, Sainani KL, Fredericson M. Electronic web-based surveys: an effective and emerging tool in research. PM R. 2010;2(4):307-9.
184. Chizawsky LL, Estabrooks CA, Sales AE. The feasibility of Web-based surveys as a data collection tool: a process evaluation. Appl Nurs Res. 2011;24(1):37-44.
185. McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, et al. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technol Assess. 2001;5(31):1-256.
186. McHorney CA, Ware JE, Jr., Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care. 1994;32(1):40-66.
187. Kundi M. Causality and the interpretation of epidemiologic evidence. Environ Health Perspect. 2006;114(7):969-74.
188. Turner S. Childhood respiratory cohort studies: do they generate useful outcomes? Breathe. 2012;8(3):194-204.
189. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg. 2010;126(6):2234-42.
190. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: consolidating data harmonization--how to obtain quality and applicability? Am J Epidemiol. 2011;174(3):261-4; author reply 5-6.
191. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BHR, Perola M, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013;10(1):12.
192. Bruin AD, Picavet HS, Nossikov A. World Health Organization. Health interview surveys: towards international harmonization of methods and instruments. 1996.
193. Rolland B, Reid S, Stelling D, Warnick G, Thornquist M, Feng Z, et al. Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach. Am J Epidemiol. 2015;182(12):1033-8.
194. Fortier I, Doiron D, Little J, Ferretti V, L'Heureux F, Stolk RP, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol. 2011;40(5):1314-28.
195. Harris JR, Burton P, Knoppers BM, Lindpaintner K, Bledsoe M, Brookes AJ, et al. Toward a roadmap in global biobanking for health. Eur J Hum Genet. 2012;20(11):1105-11.
196. Harrell FEJ. Clinical prediction models: a practical approach to development, validation, and updating. Springer. 2001.
197. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007;2:59-77.
198. Prosperi MC, Marinho S, Simpson A, Custovic A, Buchan IE. Predicting phenotypes of asthma and eczema with machine learning. BMC Med Genomics. 2014;7 Suppl 1:S7.
199. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189-202.
200. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52(10):935-42.

201. Bedogni G. Clinical Prediction Models-a Practical Approach to Development, Validation and Updating. J R Stat Soc a Stat. 2009;172:944-.
202. Savenije OEM, Kerkhof M, Koppelman GH, Postma DS. Predicting who will have asthma at school age among preschool children. Journal of Allergy and Clinical Immunology. 2012;130(2):325-31.
203. Sears MR. Predicting asthma outcomes. J Allergy Clin Immunol. 2015;136(4):829-36; quiz 37.
204. Derksen S, Keselman HJ. Backward, Forward and Stepwise Automated Subset-Selection Algorithms - Frequency of Obtaining Authentic and Noise Variables. Brit J Math Stat Psy. 1992;45:265-82.
205. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med. 2008;3:17.
206. Agresti A. Logistic regression. John Wiley & Sons, Inc. 2002.
207. Weinmayr G, Weiland SK, Bjorksten B, Brunekreef B, Buchele G, Cookson WO, et al. Atopic sensitization and the international variation of asthma symptom prevalence in children. Am J Respir Crit Care Med. 2007;176(6):565-74.
208. Yamashita T, Yamashita K, Kamimura R. A stepwise AIC method for variable selection in linear regression. Commun Stat-Theor M. 2007;36(13-16):2395-403.
209. Halken S. Prevention of allergic disease in childhood: clinical and epidemiological aspects of primary and secondary allergy prevention. Pediatr Allergy Immunol. 2004;15 Suppl 16:4-5, 9-32.
210. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc B. 2011;73:273-82.
211. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. The Journal of Machine Learning Research. 2004;5:1089-105.
212. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science. 1 Feb 1986:54-75.
213. Kyung M, Gill J, Ghosh M, Casella G. Penalized Regression, Standard Errors, and Bayesian Lassos. Bayesian Anal. 2010;5(2):369-411.
214. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. 1996 Jan 1;Series B (Methodological):267-88.
215. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. 2005;Series B (Statistical Methodology) 67(2):301-20.
216. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1-22.
217. Goeman JJ. L1 and L2 Penalized Regression Models. Technical Document for CRAN. 2010.
218. Berger JO. Statistical decision theory and Bayesian analysis. Springer Science & Business Media. 2013 Mar 14.
219. Newcombe PJ, Raza Ali H, Blows FM, Provenzano E, Pharoah PD, Caldas C, et al. Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival. Stat Methods Med Res. 2017;26(1):414-36.
220. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;82(4):711-32.
221. Newcombe PJ, Reck BH, Sun J, Platek GT, Verzilli C, Kader AK, et al. A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. Genet Epidemiol. 2012;36(1):71-83.
222. Kohn R, Smith M, Chan D. Nonparametric regression using linear combinations of basis functions. Statistics and Computing. 2001;11(4):313-22.
223. Abrams K, Ashby D, Errington D. A Bayesian approach to Weibull survival models--application to a cancer clinical trial. Lifetime Data Anal. 1996;2(2):159-74.
224. Huang A, Xu S, Cai X. Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity (Edinb). 2015;114(1):107-15.
225. Huang A, Xu S, Cai X. Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. BMC Genet. 2013;14:5.
226. Friston KJ, Litvak V, Oswal A, Razi A, Stephan KE, van Wijk BC, et al. Bayesian model reduction and empirical Bayes for group (DCM) studies. Neuroimage. 2016;128:413-31.
227. Ferro CAT, Fricker TE. A bias-corrected decomposition of the Brier score. Q J Roy Meteor Soc. 2012;138(668):1954-60.
228. Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. J Clin Epidemiol. 2007;60(5):491-501.

229. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5.
230. Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. Stat Med. 2000;19(1):61-70.
231. Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. Stat Med. 2000;19(16):2115-29.
232. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. CJEM. 2006;8(1):19-20.
233. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004;329(7458):168-9.
234. Rubin DB. Multiple imputation for nonresponse in surveys. John Wiley & Sons.
235. Schafer JL. Analysis of incomplete multivariate data. CRC press.
236. Peng CY, Harwell M, Liou SM, Ehman LH. Advances in missing data methods and implications for educational research. Real data analysis. 2006;3178.
237. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons.
238. Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet HC. Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Methodol. 2007;7:33.
239. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons. 2002.
240. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147-77.
241. Little RJA. Missing-Data Adjustments in Large Surveys - Reply. J Bus Econ Stat. 1988;6(3):300-1.
242. Carlin JB, Li N, Greenwood P, Coffey C. Tools for analyzing multiple imputed datasets. The Stata Journal. 2003;3(3):226-44.
243. De Smet F, Moreau Y, Engelen K, Timmerman D, Vergote I, De Moor B. Balancing false positives and false negatives for the detection of differential expression in malignancies. Brit J Cancer. 2004;91(6):1160-5.
244. Lu M, Stephens M. Variance adaptive shrinkage (vash): flexible empirical Bayes estimation of variances. Bioinformatics. 2016;32(22):3428-34.
245. Efron B. Empirical Bayes deconvolution estimates. Biometrika. 2016;103(1):1-20.
246. Witten I, Frank E, Hall M, Pal C. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 2016 Oct 1.
247. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. 2016;12(7):e1004977.
248. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19(8):1059-79.
249. Asher MI, Montefort S, Bjorksten B, Lai CK, Strachan DP, Weiland SK, et al. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. Lancet. 2006;368(9537):733-43.
250. Lukrafka JL, Fuchs SC, Moreira LB, Picon RV, Fischer GB, Fuchs FD. Performance of the ISAAC questionnaire to establish the prevalence of asthma in adolescents: a population-based study. J Asthma. 2010;47(2):166-9.
251. Yamada E, Vanna AT, Naspitz CK, Sole D. International Study of Asthma and Allergies in Childhood (ISAAC) - Validation of the written questionnaire (eczema component) and prevalence of atopic eczema among Brazilian children. J Investig Allergol Clin Immunol. 2002;12(1):34-41.
252. Valle SOR, Kuschnir FC, Sole D, Silva MAV, Silva RI, da Cunha AJLA. Validity and Reproducibility of the Asthma Core International Study of Asthma and Allergies in Childhood (ISAAC) Written Questionnaire Obtained by Telephone Survey. Journal of Asthma. 2012;49(4):390-4.
253. Patel SP, Jarvelin MR, Little MP. Systematic review of worldwide variations of the prevalence of wheezing symptoms in children. Environ Health-Glob. 2008;7.
254. Mohammad HR, Belgrave D, Kopec Harding K, Murray CS, Simpson A, Custovic A. Age, sex and the association between skin test responses and IgE titres with asthma. Pediatr Allergy Immunol. 2016;27(3):313-9.
255. Pearce N, Pekkanen J, Beasley R. How much asthma is really attributable to atopy? Thorax. 1999;54(3):268-72.
256. Rose G. Sick individuals and sick populations. Int J Epidemiol. 1985;14(1):32-8.
257. Rose G KK, and Marmot M. Rose's Strategy of Preventive Medicine: Oxford Scholarship Online; 2009.
258. Rockhill B, Kawachi I, Colditz GA. Individual risk prediction and population-wide disease prevention. Epidemiologic Reviews. 2000 Jan 1;22(1):176-80.
259. Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. J Thromb Haemost. 2013;11 Suppl 1:129-41.

260. Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values - Reply. Indian J Ophthalmol. 2008;56(4):341-.
261. Ehrlich RI, Du Toit D, Jordaan E, Volmink JA, Weinberg EG, Zwarenstein M. Prevalence and reliability of asthma symptoms in primary school children in Cape Town. Int J Epidemiol. 1995;24(6):1138-45.
262. Braun-Fahrlander C, Gassner M, Grize L, Minder CE, Varonier HS, Vuille JC, et al. Comparison of responses to an asthma symptom questionnaire (ISAAC core questions) completed by adolescents and their parents. SCARPOL-Team. Swiss Study on Childhood Allergy and Respiratory Symptoms with respect to Air Pollution. Pediatr Pulmonol. 1998;25(3):159-66.
263. Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee. Lancet. 1998;351(9111):1225-32.
264. Elphick HE, Sherlock P, Foxall G, Simpson EJ, Shiell NA, Primhak RA, et al. Survey of respiratory sounds in infants. Arch Dis Child. 2001;84(1):35-9.
265. Jackson DJ, Gangnon RE, Evans MD, Roberg KA, Anderson EL, Pappas TE, et al. Wheezing rhinovirus illnesses in early life predict asthma development in high-risk children. Am J Resp Crit Care. 2008;178(7):667-72.
266. Jartti T, Lehtinen P, Vuorinen T, Ruuskanen O. Bronchiolitis: age and previous wheezing episodes are linked to viral etiology and atopic characteristics. Pediatr Infect Dis J. 2009;28(4):311-7.
267. Calvo C, Garcia-Garcia ML, Blanco C, Pozo F, Flecha IC, Perez-Brena P. Role of rhinovirus in hospitalized infants with respiratory tract infections in Spain. Pediatr Infect Dis J. 2007;26(10):904-8.
268. Garcia-Garcia ML, Calvo C, Casas I, Bracamonte T, Rellan A, Gozalo F, et al. Human metapneumovirus bronchiolitis in infancy is an important risk factor for asthma at age 5. Pediatr Pulmonol. 2007;42(5):458-64.
269. Korppi M, Piippo-Savolainen E, Korhonen K, Remes S. Respiratory morbidity 20 years after RSV infection in infancy. Pediatr Pulmonol. 2004;38(2):155-60.
270. Kusel MM, de Klerk NH, Kebadze T, Vohma V, Holt PG, Johnston SL, et al. Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma. J Allergy Clin Immunol. 2007;119(5):1105-10.
271. Biedermann T, Skabytska Y, Kaesler S, Volz T. Regulation of T Cell Immunity in Atopic Dermatitis by Microbes: The Yin and Yang of Cutaneous Inflammation. Front Immunol. 2015;6:353.
272. Werfel T, Wittmann M. Regulatory role of T lymphocytes in atopic dermatitis. Chem Immunol Allergy. 2008;94:101-11.
273. Werfel T. The role of leukocytes, keratinocytes, and allergen-specific IgE in the development of atopic dermatitis. J Invest Dermatol. 2009;129(8):1878-91.

Appendix A

Figure A.1: Search on 14 July 2017 in Ovid MEDLINE(R) from 1946 to July week 2 of 2015

#	Searches	Results
1	exp asthma/	118835
2	(asthma* or wheez*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]	167279
3	1 or 2	167279
4	("predictive score*" or "prediction score*" or "predictive index*" or "prediction index*" or "predictive model*" or "prediction model*" or "predictive ability" or "prediction ability" or "predictive performance" or "prediction performance").mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]	35058
5	3 and 4	317
6	("asthma score*" or "asthma risk score*" or "asthma index*" or "asthma prediction*" or "prediction of asthma" or "asthma predictive" or "wheez* score*" or "wheez* index*" or "prediction of wheez*").mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]	477
7	5 or 6	702
8	**limit 7 to humans and "all child (0 to 18 years)"**	437

Figure A.2: Search on 14 July 2017 in Ovid EMBASE from 1947 to 11 July 2017

#	Searches	Results
1	exp asthma/	358605
2	(asthma* or wheez*).mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	448144
3	exp wheezing/	34122
4	1 or 2 or 3	455184
5	("predictive score*" or "prediction score*" or "predictive index*" or "prediction index*" or "predictive model*" or "prediction model*" or "predictive ability" or "prediction ability" or "predictive performance" or "prediction performance").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	82746
6	4 and 5	900
7	("asthma score*" or "asthma risk score*" or "asthma index*" or "asthma prediction*" or "prediction of asthma" or "asthma predictive" or "wheez* score*" or "wheez* index*" or "prediction of wheez*").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	1192
8	6 or 7	1829
9	limit 8 to human	1451
10	**limit 9 to (infant <to one year> or preschool child <1 to 6 years> or school child <7 to 12 years> or adolescent <13 to 17 years>)**	809

Figure A.3: The data extraction form used for the systematic review

Reviewer		

Article ID number		

First Author		

Year		

Type of prediction modelling studies		Prediction model without external validation or with external validation in independent data; external model validation studies with or without model updating

Design:		
Country		
Study name		
Study design / Source of data		Cohort, case-control, cross-sectional or other
Recruitment of study population		
Inclusion/exclusion criteria		
Ethnicity		
Recruitment year		
Sample size at the start of study		
If cohort study:		
Type of follow-up (prospective, retrospective)		

Total length of follow-up		Specify number of years/months

Description of participants' characteristics		
Participants' characteristics reported		Give list (e.g. weight, parental history of asthma, etc.) or specify table and page where list can be found
If cohort study:		
Participants' characteristics reported at baseline		As above
Participants' characteristics reported at follow-up		As above

Outcome - Asthma		
Definition of asthma		Give definition of the outcome
Asthma assessment blinded for exposure to predictors		Yes/No

Candidate predictors (predictors to be studied for their predictive performance)		
Number of candidate predictors		Give number of candidate predictors
List of candidate predictors, with age of exposure		Give a list with age, or specify table
Definition/method of measurement of candidate predictors		Give definition/method for each predictor, or specify table
Handling of predictors in the modelling (e.g. continuous, transformed)		Give information for each predictor, or specify table
Exposure assessment blinded for asthma status		Yes/No

Missing data		Yes/No

Number of participants with any missing values (include predictors and outcomes)		Specify page number or table
Number of participants with missing data for each predictors		Specify page number or table
Handling of missing data (e.g. complete-case analysis, imputation or other methods)		Specify method

Model development		
Prediction Modelling method (e.g. logistic, survival, neural networks, or machine learning techniques)		
Modelling assumptions satisfied		Reported/Not reported
Method for selection of predictors for inclusion in initial multivariable model (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)		
Method for selection of predictors final multivariable model (e.g., full model approach, backward or forward selection) and criteria used (e.g., *p* value, Akaike Information Criterion)		
Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)		Reported (if yes, specify)/Not reported

Model performance at outcome assessment 1		
Age at outcome assessment 1		
Number of participants at this assessment		
Number of asthmatics at this assessment		
Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (*c*-		Yes/No

statistic, D-statistic, log-rank) measures with confidence intervals		
Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a priori cut points were used		Yes/No

Model evaluation		
Internal validation (bootstrap, cross-validation, none)		Yes/No - Specify method (e.g. leave-one-out cross-validation) and table
External validation (temporal, geographical, different investigators, different settings)		Yes/No - Specify method and table
In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)		Yes/No - Specify method and table

Results		
Predictors identified, with age of exposure		Give a list of predictors identified and specify age of exposure
Effect size, with 95%CI, for each predictor identified (e.g. predictor weights or regression coefficients)		Specify table or text
Sensitivity (specify whether cut points were used)		Specify table or text
Specificity (specify whether cut points were used)		Specify table or text
Positive Predictive Value (specify whether cut points were used)		Specify table or text
Negative Predictive Value (specify whether cut points were used)		Specify table or text
Measure(s) of calibration (calibration plot, calibration slope, Hosmer-Lemeshow)		Specify table or text

Measure(s) of discrimination (AUC or C-statistic, log-rank, D-statistic)		Specify table or text
Measure(s) of overall performance (Brier)		Specify table or text
Any alternative presentation of the final prediction models, e.g., sum score, score chart, predictions for specific risk subgroups with performance		Specify table or text
Comparison of the distribution of predictors (including missing data) for development and validation datasets		Specify table or text

Interpretation and Discussion		
Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)		
Comparison with other studies, discussion of generalisability, strengths and limitations		

Additional references		Check all references in the paper that might be relevant for systematic review and cross-ref with existing papers included in systematic review. If not included, mention which ref from the current paper should be included.

Table A.1: Odds Ratios estimated for predictors included into predictive models of development studies included into the systematic review

Candidate predictor	**Castro-Rodriguez 2000 (API)**	**Caudri 2009 (PIAMA)**	**Pescatore 2014**
	OR (95%CI)	**OR (95%CI)**	**OR (95%CI)**
Sex (Male)		1.7 (1.3,2.3)	1.48 (NR)
Age			1.19 (NR)
Post-term delivery		2.3 (1.3,4.0)	
Low/medium Parental education		1.6 (1.1,2.3)	
Parental inhalation medication			

Wheezing frequency*	NR		
Never		Reference	
1-3 times per year		1.6 (1.1,2.3)	
≥4 times per year		2.8 (1.9,4.2)	1.65 (NR)†
Wheezing apart from colds	NR	2.3 (1.3,4.1)	1.4 (NR)
Infections			
Never		Reference	
1-2 times per year		1.7 (1.3,2.2)	
≥3 times per year		2.2 (1.4,3.3)	
Eczema	NR	2.6 (1.9,3.5)	1.52 (NR)
Parental asthma (any of the two parents)	NR	2.4 (1.8,3.3)	1.23 (NR)
Parental asthma (both parents)			1.26 (NR)
Allergic rhinitis	NR		
Eosinophilia (≥4%)	NR		
Activity disturbance*			
Any			1.28 (NR)
Moderate			1.16 (NR)
A lot			1.63 (NR)
Shortness of breath*			
Sometimes			1.98 (NR)
Always			1.56 (NR)
Exercise-related wheeze/cough			1.26 (NR)
Aeroallergen-related wheeze/cough			1.22 (NR)

*No reference category for these predictors in the Leicester study. †Wheezing frequency is estimated for 3≥times per year. Abbreviations – OR: odd ratio; CI: confidence intervals; NR: not reported.

tistical methods used among studies

	Model development	Prediction rule	Missing data	Model evaluation	External validation study
			No modelling		
ez	Existing knowledge to select significant predictors; Chi-square test to compare proportions	Stringent API and Loose API	complete case *Proportion of missing:* 20%	None	*Devulapalli 2007; Rodriguez-Martinez 2011; Leonardi 2011*
			Logistic regression		
	Existing knowledge, univariate analysis (P value < 0.05) and stepwise method (P value < 0.01); Multivariable logistic model	PIAMA	complete case analysis and imputation using MICE *Proportion of missing:* predictors 15%; outcomes 12%. In total each predictor has less than 2%	Internally validated by bootstrap sampling method (200 bootstrap samples)	*Rodriguez-Martinez 2011; Hafkamp-de Groen 2013*
			Advanced methods		
4	Existing knowledge and only those that were available in primary care and did not require repeated observations were selected; LASSO regression	Risk score	NR *Proportion of missing:* in total each predictor has less than 5.8%	Internally validated by leave-one-out cross-validation technique	*Grabenhenrich 2014*

Table A.3: Performance measures among studies and target populations

Source of study	Target population	ID	Prediction rule	Outcome age	%Se (%) (95% CI)	%Sp (95% CI)	%PPV (95% CI)	%NPV (95% CI)	LR (+)	LR (-)	Others
Development	*General*	*Castro-Rodriguez J. A. 2000*	Stringent API	6	28 (24–30)	96 (95–98)	48 (44–51)	92 (90–93)	7.43*	0.75*	NR
				11	15 (13–17)	96 (95–97)	42 (39–45)	86 (83–88)	3.85*	0.88*	
			Loose API	6	57 (53–60)	81 (78–83)	26 (24–29)	94 (92–95)	2.92*	0.54*	
				11	40 (37–43)	80 (77–82)	27 (24–30)	88 (85–90)	1.97*	0.75*	
	High-risk	*Caudri 2009*	PIAMA	7-8	60 (NR)	76 (NR)	23 (NR)	94 (NR)	2.5*	0.53*	AUC=0
		Pescatore 2014	Risk score	6-8	72 (NR)	71 (NR)	49 (NR)	86 (NR)	2.5	0.39	Brier = AUC =
Validation	*General*	*Leonardi 2011*	Stringent API	7	37 (30-44)	93 (92-95)	40 (32-48)	93 (91-94)	5.29*	0.68*	Brier=1 AUC=0
				10	32 (24-41)	94 (92-95)	35 (27-45)	92 (91-94)	5.33*	0.72*	Brier=1 AUC=0
			Loose API	7	57 (50-64)	80 (78-82)	26 (21-30)	94 (92-95)	2.85*	0.54*	Brier=1 AUC=0
				10	57 (48-66)	81 (78-83)	25 (20-31)	94 (93-96)	3*	0.53*	Brier=1 AUC=0
	High-risk	*Devulapalli 2008*	Stringent API	10	57 (47–66)	83 (79–87)	48 (39–57)	87 (83–91)	3.34*	0.52*	NR
			Loose API		60 (50-69)	79 (74-83)	44 (36-53)	88 (84–91)	2.85*	0.51*	
		Rodriguez-Martinez 2011	Stringent API	5-6	43 (25–64)	79 (68–87)	38 (21–57)	83 (72–90)	2.06	0.72	NR
			Loose API		71 (50–86)	33 (24–45)	24 (15–36)	80 (63–91)	1.07	0.86	
			PIAMA		55 (43–66)	79 (67–88)	75 (61–85)	60 (49–70)	2.59	0.58	
		Hafkamp-de Groen 2013	PIAMA	6	NR	NR	NR	NR	NR	NR	AUC=0
		Grabenhenrich 2014	Pescatore's score	8	82 (NR)	69 (NR)	40 (NR)	94 (NR)	2.63	0.26	Brier= AUC=0

*Calculated using measures available. Abbreviations - CI: confidence intervals; Se: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC: area under the (ROC) curve; NR: not reported.

Table A.4: Critical appraisal of the eight selected studies based on the CHARMS checklist
Abbreviations - L: low; M: medium; H: high.

ID	Risk of bias				
	Participant selection	**Outcome**	**Predictors**	**Sample flow/attrition**	**Analysis**
Castro-Rodriguez 2000	L	L	L	M	M
Caudri 2009	L	L	L	M	L
Pescatore 2014	L	L	L	M	L
Leonardi 2011	L	L	L	M	L
Devulapalli 2008	L	L	L	M	M
Rodriguez-Martinez 2011	M	L	L	H	M
Hafkamp-de Groen 2013	L	L	L	M	M
Grabenhenrich 2014	L	L	L	M	L

Appendix B

Figure B.1: The Ashford survey

Information Sheet and Consent Form

Background:

"Asthma in Ashford" is a research project that started nearly 20 years ago. You may (or may not!) remember taking part in it before; the last time we saw you was about 5 years ago. The study is about why some children and young people get asthma and allergies while others do not. So far the information you have provided so far has been really useful and we're very grateful for your help. For example we learned that while lots of young children get inhalers, most of them don't go on to get real asthma; and that getting antibiotics when you are young doesn't increase the chance that you will get asthma.

Now we're asking you for some more help; we are trying to understand more about why some young people continue to have asthma while others grow out of it. Please note that we'd like your help *even if you don't have asthma or any allergies*.

What we want you to do:

We would like you to fill in a short questionnaire about what you are doing now and about any asthma or allergies you might have; this should take only about 10 minutes. Remember, we would like your help *even if you don't have asthma or any allergies*. You don't have to take part but if you agree you will be ask to sign a consent form (below).

You can complete the questionnaire on paper and send it back to us using the stamped envelope we have sent you. Or you can do it online through Imperial College's website; just copy the URL «Short_Link» into your browser and it will take you to the questionnaire; you will be asked to give an **ID number** which is: **«STUDYNO»**.

If you complete the questionnaire and send it back
we will enter your name in a lottery with 5 prizes of £50!

What will happen to the information you give us?

All your answers to the questions will be kept confidential and will only be reported in a way that cannot identify you. Nobody outside our research team will have access to your answers. They will be stored in a secure database until we have finished analysing them; then they will be destroyed.

Is there any risk to me if I take part?

We are not testing any new medicines or treatments and so there are no risks to your health from taking part. Moreover, the questions we are asking you are not tricky or sensitive ones.

Is there any benefit to me if I take part?

There are no direct benefits to you if you take part but we hope, with your help, to learn more about asthma and improve treatment for children and other young people in the future.

Who has reviewed the study?

Research studies such as this are governed by research guidelines and legislation. This study has been reviewed and approved by Imperial College Research Ethics Committee.

I confirm I have received, read and understood the "Asthma in Ashford" sheet (above) ☐

I understand that my participation is voluntary and I am free to withdraw at any time without giving a reason ☐

I agree to complete a short questionnaire about my health condition. ☐

Please tick here if you wish to opt out (and then send this back to us) ☐

Welcome back to Asthma in Ashford!

Part A. YOURSELF

A1. Please write your full name here: ☐

A2. Are you still in education?

☐ Yes, full-time education

☐ Yes, part-time education

☐ No, I am not in any education → ***if No, go to A3 on this page***

A2a. Which of the following are you currently studying for?

☐ A levels or AS levels

☐ NVQ or similar

☐ University

☐ Other (please tick & describe) _____

A3. At the moment do you have a paid job?

☐ Yes, full-time paid work (30 or more hours a week)

☐ Yes, part-time paid work (less than 30 hours a week)

☐ No, I am not working in a paid job → ***if No, go to A4 on this page***

A3a. What is the job title of your current paid job?

_____

_____

_____

A3b. Please describe the main things you do in your current paid job:

_____

_____

A4. If you are not in education and are not in paid work, are you:

- ☐ Unemployed and looking for a job
- ☐ Long-term sick or disabled
- ☐ Looking after home/family
- ☐ Other (please tick & describe) _____

A5. Do you have children of your own?

- ☐ Yes
- ☐ No → ***if No, go to part B on this page***

A5a. How many children do you have? ☐

Part B: YOUR HOME

B1. What is your full post code? ☐☐☐☐ ☐☐☐

B2. Who are you living with?

- ☐ I live on my own
- ☐ I live with my family
- ☐ I live with my friend(s) / flatmate(s)
- ☐ I live with my partner
- ☐ Other (please tick & describe) _____

B3. Including yourself, how many people live in your household now? ☐

B4. Including yourself, how many people in your household are 18 years or older? ☐

B5. How many years have you lived in the home where you live now? (If you can't remember, make a guess and tick the guess box as well)

Number of years:	Tick if this is a guess:
	☐

B6. Is there any damp or condensation in your home?

- ☐ Yes → *if Yes, go to B6a. on this page*
- ☐ No
- ☐ Don't know

B6a. How much of a problem is the damp or condensation?

- ☐ Not a problem at all
- ☐ Not serious
- ☐ Fairly serious
- ☐ Very serious

B7. Is there any mould in your home?

- ☐ Yes → *if Yes, go to B7a. on the next page*
- ☐ No
- ☐ Don't know

B7a. How much of a problem is the mould?

- ☐ Not a problem at all
- ☐ Not serious
- ☐ Fairly serious
- ☐ Very serious

B8. How heavy is the traffic on the road where you live?

- ☐ Hardly any traffic
- ☐ Not very heavy
- ☐ Quite heavy
- ☐ Very heavy

B9. How often do lorries or buses pass your home?

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often

B10. To heat your home in winter what methods do you use? (Please tick <u>all</u> boxes that apply)

- ☐ Central heating or storage heaters
- ☐ Gas fires (mains or bottled gas)
- ☐ Electricity
- ☐ Wood stoves or wood fires
- ☐ Coal fires
- ☐ Paraffin heaters
- ☐ Other (please tick & describe) _____

B11. What kind of <u>cooker</u> do you have at home?

- ☐ Gas
- ☐ Electric
- ☐ Other (please tick & describe) _____

B11a. If your cooker uses gas, is this:

- ☐ Only the rings of the cooker
- ☐ Only the oven of the cooker
- ☐ Both the rings and the oven

B12. <u>Excluding yourself</u>, how many people living in your home are smokers? ☐

B13. Excluding yourself, who is or are the smokers in your home? (Please tick all that apply)

- ☐ Mother
- ☐ Father
- ☐ Brother / Sister
- ☐ Partner
- ☐ Friend(s) / Flatmate(s)
- ☐ Nobody → *if Nobody, go to B14 on this page*
- ☐ Other (please tick & describe) _____

B13a. Do you have a rule that smoking never happens in particular rooms?

- ☐ No smoking inside is allowed
- ☐ Smoking is only allowed in some rooms
- ☐ Smoking is allowed anywhere

B14. Have you got any pets at your home?

- ☐ Yes
- ☐ No → *if No, go to part C on this page*

B14a. Please tell us how many of each type you have:

a. Cats ☐
b. Dogs ☐
c. Birds ☐
d. Rabbits or rodents (e.g. guinea pigs, hamsters, mice) ☐
e. Other (please describe and specify) ☐

Part C: WHEEZING AND ASTHMA

C1. Have you had wheezing or whistling in your chest in the past 12 months?

- ☐ Yes
- ☐ No → *if No, go to part C2 on page 7*

C1a. How many attacks of wheezing have you had in the past 12 months?

- ☐ None
- ☐ 1-3
- ☐ 4-12
- ☐ More than 12

C1b. In the past 12 months, how often, on average, has your sleep been disturbed due to wheezing?

- ☐ Never woken with wheezing
- ☐ Less than one night per week
- ☐ One or more nights per week

C1c. In the past 12 months, has your chest sounded wheezy during or after exercise?

- ☐ Yes
- ☐ No

C1d. In the past 12 months, has wheezing ever been severe enough to limit your speech to only one or two words at a time between breaths?

- ☐ Yes
- ☐ No

C1e. How many days of school/college or work (paid employment) have you missed in the past 12 months due to wheezing in the chest? (If you can't remember, make a guess and tick the guess box as well)

	Please specify	Tick if this is a guess
Number of days off school/college:		☐
Number of days off work:		☐

C1f. During the last 12 months, has the wheezing in your chest been better at the weekends?

- ☐ Yes
- ☐ No

C1g. Has the wheezing in your chest been better when away from work for a period longer than a weekend? If you don't have a job, tick 'No'

- ☐ Yes
- ☐ No

C2. Have you ever had asthma?

- ☐ Yes
- ☐ I am not sure
- ☐ No → *if No and if you also said No to the question about wheeze (C1 on page 5) then go to part E on page 8; otherwise go to part D on this page*

Part D: TREATMENTS AND TRIGGERS

D1. Is there anything that you know that sets off your wheeze or asthma?

- ☐ Yes
- ☐ No → *if No, go to D2 on this page*

D1a. Which of these makes you wheeze? (Please tick all that apply)

- ☐ Exercising
- ☐ Infections
- ☐ Pollen
- ☐ Animals
- ☐ House dust
- ☐ Stress
- ☐ Other (please tick & describe)

_____

D2. Do you currently take any regular treatment for your wheezing (or asthma)?

- ☐ Yes
- ☐ No → *if No, go to D3 on this page*

D2a. Please write down the name of the current treatments that you use:

_____

_____

_____

D2b. Does it help?

☐ Yes

☐ No

D2c. How many courses of steroid tablets have had for your wheezing or asthma in the last 12 months? ☐

D3. In the last 12 months have you been to casualty (A/E) because of wheeze or asthma?

☐ Yes

☐ No

D4. In the last 12 months have you been kept in hospital because of wheeze or asthma?

☐ Yes

☐ No

Part E: COUGH, HAYFEVER, ECZEMA & RASH

E1. In the past 12 months, have you had a dry cough at night, apart from a cough associated with a cold or chest infection?

☐ Yes

☐ No

E2. Have you ever had hayfever?

☐ Yes

☐ No

E3. In the past 12 months, have you had a problem with sneezing or a runny or blocked nose when you *DID NOT* have a cold or flu (such as running a temperature or feeling generally unwell)?

☐ Yes

☐ No → ***if No, go to E4 on this page***

Note: these questions are about when you DO NOT have a cold or the flu

E3a. In the past 12 months, has this nose problem been accompanied by itchy or watery eyes?

☐ Yes

☐ No

E3b. In which of the past 12 months, did this nose problem occur? (Please tick all that apply)

☐ January	☐ February	☐ March	☐ April
☐ May	☐ June	☐ July	☐ August
☐ September	☐ October	☐ November	☐ December

E4. Have you ever had eczema?

☐ Yes

☐ No

E5. Have you had an itchy red rash at any time in the past 12 months?

☐ Yes

☐ No → *if No, go to E6 on this page*

E5a. Has this itchy rash at any time affected any of the following places: the folds of the elbows, behind the knees, in front of the ankles, under the buttocks, or around the neck, ears or eyes?

☐ Yes

☐ No

E6. Have you ever suffered from food allergy?

☐ Yes

☐ No → *if No, go to part F on the next page*

E6a. Which food are you allergic to? (Please tick all that apply)

☐ Eggs

☐ Wheat

☐ Cow's milk

☐ Peanuts, nuts or sesame seeds

☐ Other (please tick & describe) _____

Part F: YOUR HEALTH

F1. In general, would you say your health is:

- ☐ Excellent
- ☐ Very good
- ☐ Good
- ☐ Fair
- ☐ Poor

F2. Compared to 5 years ago, how would you rate your health in general now?

- ☐ Much better
- ☐ Much worse
- ☐ Hasn't really changed
- ☐ A bit better
- ☐ A bit worse

F3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

	Yes, limited a lot	Yes, a little	No, not limited at all
Vigorous activities (e.g. running, lifting heavy objects, participating in strenuous sports)	☐	☐	☐
Climbing several flight of stairs	☐	☐	☐
Climbing one flights of stairs	☐	☐	☐
Walking more than a mile	☐	☐	☐
Walking 100 meters	☐	☐	☐

F4. During the past 30 days, how much of the time has your physical health interfered with your social activities (e.g. meeting friends or going out)?

- ☐ None of the time
- ☐ A little of the time
- ☐ Sometimes
- ☐ All of the time

F5. How TRUE or FALSE is <u>each</u> of the following statements for you?

	Definitely True	Mostly True	Mostly False	Definitely False	Don't know
I seem to get ill more easily than other people	☐	☐	☐	☐	☐
I expect my health to get worse	☐	☐	☐	☐	☐
I am as healthy as anybody I know	☐	☐	☐	☐	☐

F6. Do you have any long-standing illness or disability? (By long-standing we mean anything that has troubled you over a period of time or that is likely to affect you over a period of time.)

☐ Yes

☐ No → ***if No, go to part G on this page***

F6a. Please describe your illness or disability:

_____

F6b. Does your illness or disability limit your activities in any way?

☐ Yes

☐ No

Part G: EXERCISE

G1. During the <u>past year</u>, how often did you do any exercise (going to the gym, brisk walking or any sports activity)?

☐ Never

☐ Once or twice

☐ Less than monthly

☐ Monthly

☐ Weekly

☐ Daily or almost daily

G2. Does anything stop you exercising?

☐ Yes

☐ No → ***if No, go to part I on the next page***

G2a. What stops you exercising?

_____

_____

_____

Part I: SMOKING

I1. Have you ever smoked a whole cigarette (including roll-ups)?

☐ Yes

☐ No → *if No, go to part J on this page*

I1a. How old were you when you first smoked a whole cigarette? ☐

I1b. How many cigarette have you smoked altogether in your life?

☐ Less than 5

☐ 5-19

☐ 20-49

☐ 50-99

☐ 100 plus

I1c. Have you smoked any cigarette in the past 30 days?

☐ Yes

☐ No → *if No, go to part J on this page*

I1d. Do you smoke every day? ☐ Yes ☐ No

Part J: COMPLETION

J1. Did you complete the questionnaire all on your own?

☐ Yes → *if Yes, go to J2 on next page*

☐ No

J1a. Who helped you to complete the questionnaire? (Please tick all that apply)

☐ Mother ☐ Father

☐ Other (please tick & describe) _____

J1b. Does she/he live in the same house as you?

☐ Yes

☐ No

J2. Please give <u>your</u> date of birth: ☐☐/☐☐/☐☐☐☐

J3. Please write <u>your</u> email address here: ☐

Space for any additional comments you would like to make:

Thank you very much for your help!

Send the questionnaire back to us to get an entry into our lottery!

Figure B.2: Sample letter to possible participants

Date

Dear *name of child*,

We hope this finds you well.

Do you remember the **Asthma in Ashford** study? You (and your mother) were kind enough to help us out for many years; we last saw you in 2008.

We would like to find out how you are getting on. To do this we are asking you to fill in a short survey which asks about the same sort of things we asked when you were younger; questions about you and your health and what you are up to (it doesn't ask any sensitive questions!). **As before, we would like ALL of you to do this, whether or not you have asthma.**

You can do the survey in one of two ways:

1. **Online**; here is a web link: https://imperial.eu.qualtrics.com/SE/?Q_SS=czJVpoFNKSatWCN_d1e0l63uPlV6FBb&_=1

or

2. By hand; with this letter there is a copy of the questionnaire and a stamped envelope for you to send it back to us.

You don't have to do this if you don't want to; but as an encouragement, if you do we'll put your name into a small lottery (with **five prizes of £50**!).

This survey has full ethical approval (ref). As before, all the information we collect and use in the study is stored and used in complete confidence.

With best wishes,

Ms Sylvia Colicino and Professor Paul Cullinan

Ref: Joint Research Compliance Office (JRCO)

Appendix C

Table C.1: Proportion of missing values among the STELAR cohorts and candidate predictors collected by age five years in the high-risk population

Candidate predictors	ALSPAC	Ashford	IoW	MAAS	SEATON
	(%)	(%)	(%)	(%)	(%)
Sex	0.00	0.00	0.00	0.00	0.00
SES	4.21	0.00	37.32	1.23	0.00
Maternal asthma	0.84	0.00	0.47	0.00	0.00
Maternal eczema	0.84	0.00	0.47	0.00	0.00
Maternal hayfever	0.84	0.00	0.00	0.00	0.00
Paternal asthma	8.63	0.00	0.00	0.00	0.00
Paternal eczema	9.47	0.00	0.00	0.00	0.00
Older siblings	35.16	0.00	9.57	0.00	0.00
No. of older siblings	35.16	0.00	9.57	0.00	0.00
Birth order	35.16	0.00	0.00	0.00	0.00
Maternal age at recruitment	1.26	0.00	0.00	0.00	0.00
Premature birth	0.84	0.00	0.96	0.00	4.35
Ever breast fed	0.63	0.00	0.48	0.00	0.00
Vaginal delivery	1.68	0.00	66.03	0.00	5.43
Age weaning (weeks)	3.16	0.00	3.83	0.00	4.35
Cat at 1 yr	2.11	0.00	44.5	0.00	0.00
Cat at 4/5 yrs	3.16	0.00	0.00	1.23	0.00
Dog at 1 yr	2.11	0.00	44.5	2.06	0.00
Dog at 4/5 yrs	3.16	0.00	0.00	1.23	0.00
Moved house by 2 yrs	5.26	0.00	0.00	0.00	4.35
Maternal smoking during pregnancy	6.32	0.00	0.48	0.00	0.00
Maternal smoking at 1 yr	1.05	0.00	2.87	0.00	5.43
Current wheeze at 1 yr	0.00	0.00	0.00	0.00	0.00
Current wheeze at 2/3 yrs	0.00	0.00	0.00	0.00	0.00
Current wheeze at 4/5 yrs	0.00	0.00	0.00	0.00	0.00
Frequency of wheezing at 1 yrs	0.21	0.00	6.37	0.00	5.53
Frequency of wheezing at 2/3 yrs	5.68	0.00	0.00	0.00	0.00
Frequency of wheezing at 4/5 yrs	2.11	0.00	0.00	0.00	5.53
Eczema at 1 yr	1.68	0.00	7.17	0.00	0.00
Eczema at 2/3 yrs	4.42	0.00	0.40	0.00	0.00
Eczema at 4/5 yrs	3.37	0.00	0.00	0.00	0.00

Abbreviations - SES: socioeconomic status; yr(s): year(s)

Table C.2: Univariate analysis in the development data (ALSPAC and Ashford study) for the high-risk population

Candidate Predictor	Age 15–20		
	Non-asthma	Asthma	
	n (%) or median (IQR)	n (%) or median (IQR)	*P*-value[a]
Sex (Female)	200 (50.6)	63 (45.3)	0.328
SES			0.750
Professional/managerial occupations	116 (30.6)	41 (30.4)	
Skilled non-manual occupations	96 (25.3)	40 (29.6)	
Skilled manual occupations	117 (30.9)	35 (25.9)	
Partly-skilled occupations	42 (11.1)	17 (12.6)	
Unskilled occupations	8 (2.1)	2 (1.5)	
Maternal asthma	66 (16.8)	36 (26.3)	**0.022**
Maternal eczema	130 (33.1)	48 (35.0)	0.754
Maternal hayfever	140 (35.6)	65 (47.4)	**0.019**
Paternal asthma	55 (15.1)	25 (19.4)	0.322
Paternal eczema	31 (8.6)	17 (13.2)	0.186
Older siblings	188 (69.6)	69 (71.1)	0.882
No. of older siblings	2 (2.0)	2 (1.0)	0.083
Birth order			0.886
first-born	82 (30.4)	28 (28.9)	
second-born	123 (45.6)	47 (48.4)	
≥3	65 (24.0)	22 (22.7)	
Maternal age at recruitment	29 (6.0)	30 (5.3)	0.510
Premature birth	25 (9.1)	7 (7.6)	0.824
Ever breast fed	314 (79.9)	118 (85.5)	0.184
Vaginal delivery	224 (82.1)	73 (81.1)	0.966
Age weaning (weeks)	13.04 (4.3)	13.04 (4.3)	0.244
Cat at 1 yr	166 (42.6)	45 (33.6)	0.084
Cat at 4/5 yrs	154 (40.0)	47 (35.1)	0.365
Dog at 1 yr	142 (36.4)	32 (23.9)	**0.011**
Dog at 4/5 yrs	148 (38.4)	41 (30.6)	0.128
Moved house by 2 yrs	72 (18.9)	24 (18.6)	>0.99
Maternal smoking during pregnancy	63 (17.1)	23 (17.0)	>0.99
Maternal smoking at 1 yr	157 (40.2)	64 (46.4)	0.240
Current wheeze at 1 yr	196 (49.6)	71 (51.1)	0.844
Current wheeze at 2/3 yrs	202 (51.1)	82 (59.0)	0.134
Current wheeze at 4/5 yrs	360 (89.8)	139 (97.9)	**0.004**
Frequency of wheezing at 1 yr			0.449
Never	209 (52.9)	74 (53.6)	
Sometimes	132 (33.4)	40 (29.0)	
Often	54 (13.7)	24 (17.4)	
Frequency of wheezing at 2/3 yrs			0.208
Never	180 (48.1)	54 (40.7)	
Sometimes	141 (37.7)	53 (39.8)	
Often	53 (14.2)	26 (19.5)	
Frequency of wheezing at 4/5 yrs			**<0.001**
Never	41 (10.5)	3 (2.2)	
Sometimes	248 (63.8)	68 (50.4)	
Often	100 (25.7)	64 (47.4)	
Eczema at 1 yr	153 (38.7)	80 (57.1)	**<0.001**
Eczema at 2/3 yrs	146 (38.6)	81 (60.0)	**<0.001**
Eczema at 4/5 yrs	155 (40.6)	87 (64.0)	**<0.001**

Abbreviation - SES: socioeconomic status; yr(s): years

[a] *P*-values obtained by Chi-square test or Fisher's exact test. In bold, statistically significant *p*-values ($p < 0.05$).

Appendix D

Table D.1: Proportion of missing values among the STELAR cohorts and candidate predictors collected by age 10 years in the general population

Candidate predictor	ALSPAC (%)	Ashford (%)	IoW (%)	MAAS (%)	SEATON (%)
Sex	0.00	0.00	0.00	0.00	0.00
SES	5.78	0.83	50.67	2.93	1.13
Maternal asthma	2.68	0.00	1.12	0.00	0.00
Maternal eczema	2.68	0.00	0.67	0.00	0.00
Maternal hayfever	2.68	0.00	0.67	0.00	0.00
Paternal asthma	12.34	0.00	1.35	0.00	0.00
Paternal eczema	12.63	0.00	1.35	0.00	0.00
Older siblings	40.54	0.00	6.73	2.53	0.00
No. of older siblings	40.54	0.00	6.73	0.93	0.00
Birth order	40.54	0.00	17.26	0.93	0.00
Maternal age at recruitment	1.58	0.00	0.00	4.52	0.00
Premature birth	0.46	0.00	1.57	9.04	3.53
Ever breast fed	1.50	0.00	10.76	5.05	4.65
Vaginal delivery	1.29	0.00	67.49	14.23	6.06
Age weaning (weeks)	5.19	0.00	14.80	55.45	7.90
Cat at 1 yr	6.03	0.00	51.79	1.73	4.94
Cat at 4/5 yrs	10.76	0.00	17.26	2.53	4.09
Dog at 1 yr	6.03	0.00	51.79	1.60	4.94
Dog at 4/5 yrs	10.76	0.00	17.26	2.93	4.09
Moved house by 2 yrs	10.26	0.00	17.26	0.00	8.46
Maternal smoking during pregnancy	7.85	0.00	0.67	0.00	0.00
Maternal smoking at 1 yr	1.52	0.00	12.56	0.00	8.60
Current wheeze at 1 yr	2.26	0.00	12.56	55.45	4.94
Current wheeze at 2/3 yrs	7.83	0.00	25.11	5.59	12.27
Current wheeze at 4/5 yrs	11.61	0.00	17.26	2.53	15.94
Current wheeze at 8 or 10 yrs	13.63	7.63	4.60	2.11	24.31
Frequency of wheezing at 1 yrs	2.35	0.00	19.73	55.45	6.77
Frequency of wheezing at 2/3 yrs	8.44	0.00	25.11	5.59	12.27
Frequency of wheezing at 4/5 yrs	11.97	0.00	17.26	2.53	17.07
Frequency of wheezing at 8 or 10 yrs	14.15	7.63	4.60	2.11	24.78
Eczema at 1 yr	5.38	0.00	21.30	55.45	4.94
Eczema at 2/3 yrs	8.23	0.00	19.51	5.59	12.27
Eczema at 4/5 yrs	10.34	0.00	17.26	2.53	15.94
DD asthma by 4/5 or 7.5 yrs	12.65	0.00	17.26	2.53	16.36
Asthma treatment at 4/5 or 6.4 yrs	13.34	0.00	17.26	2.53	16.22
SPT cat at 4/5 or 7 yrs	27.86	6.20	30.94	9.97	45.13
SPT house dust mite at 4/5 or 7 yrs	27.32	5.79	30.94	10.11	45.13
SPT mixed grass at 4/5 or 7 yrs	27.59	6.20	30.94	10.11	45.13

Abbreviation - SES: socioeconomic status; yr(s): years; DD: doctor's diagnosis; SPT: skin prick testing

Table D.2: Proportion of missing values among the five STELAR cohorts and candidate predictors collected by age 10 years in the high-risk population

Candidate predictor	ALSPAC	Ashford	IoW	MAAS	SEATON
	(%)	(%)	(%)	(%)	(%)
Sex	0.00	0.00	0.00	0.00	0.00
SES	4.21	0.00	37.32	1.23	0.00
Maternal asthma	0.84	0.00	0.47	0.00	0.00
Maternal eczema	0.84	0.00	0.47	0.00	0.00
Maternal hayfever	0.84	0.00	0.00	0.00	0.00
Paternal asthma	8.63	0.00	0.00	0.00	0.00
Paternal eczema	9.47	0.00	0.00	0.00	0.00
Older siblings	35.16	0.00	9.57	0.00	0.00
No. of older siblings	35.16	0.00	9.57	0.00	0.00
Birth order	35.16	0.00	0.00	0.00	0.00
Maternal age at recruitment	1.26	0.00	0.00	0.00	0.00
Premature birth	0.84	0.00	0.96	0.00	4.35
Ever breast fed	0.63	0.00	0.48	0.00	0.00
Vaginal delivery	1.68	0.00	66.03	0.00	5.43
Age weaning (weeks)	3.16	0.00	3.83	0.00	4.35
Cat at 1 yr	2.11	0.00	44.5	0.00	0.00
Cat at 4/5 yrs	3.16	0.00	0.00	1.23	0.00
Dog at 1 yr	2.11	0.00	44.5	2.06	0.00
Dog at 4/5 yrs	3.16	0.00	0.00	1.23	0.00
Moved house by 2 yrs	5.26	0.00	0.00	0.00	4.35
Maternal smoking during pregnancy	6.32	0.00	0.48	0.00	0.00
Maternal smoking at 1 yr	1.05	0.00	2.87	0.00	5.43
Current wheeze at 1 yr	0.00	0.00	0.00	0.00	0.00
Current wheeze at 2/3 yrs	0.00	0.00	0.00	0.00	0.00
Current wheeze at 4/5 yrs	0.00	0.00	0.00	0.00	0.00
Current wheeze at 8 or 10 yrs	3.23	0.00	1.57	4.72	0.00
Frequency of wheezing at 1 yrs	0.21	0.00	6.37	0.00	5.53
Frequency of wheezing at 2/3 yrs	5.68	0.00	0.00	0.00	0.00
Frequency of wheezing at 4/5 yrs	2.11	0.00	0.00	0.00	5.53
Frequency of wheezing at 8 or 10 yrs	3.23	0.00	1.57	4.72	0.00
Eczema at 1 yr	1.68	0.00	7.17	0.00	0.00
Eczema at 2/3 yrs	4.42	0.00	0.40	0.00	0.00
Eczema at 4/5 yrs	3.37	0.00	0.00	0.00	0.00
DD asthma by 4/5 or 7.5 yrs	3.37	0.00	0.00	0.00	0.00
Asthma treatment at 4/5 or 6.4 yrs	6.32	0.00	0.00	0.00	1.01
SPT cat at 4/5 or 7 yrs	24.42	8.47	8.63	11.30	36.68
SPT house dust mite at 4/5 or 7 yrs	23.58	6.78	8.63	11.30	36.68
SPT mixed grass at 4/5 or 7 yrs	23.58	8.47	8.63	11.30	36.68

Abbreviation - SES: socioeconomic status; yr(s): years; DD: doctor's diagnosis; SPT: skin prick testing

Table D.3: Univariate analysis in the development data (ALSPAC and Ashford study) for the general population

Candidate predictor	Age 15–20		
	Non-asthma	Asthma	
	n (%) or median (IQR)	n (%) or median (IQR)	*P* value[a]
Sex (Female)	1923 (41.3)	154 (38.8)	0.354
SES			0.276
Professional/managerial occupations	1417 (32.2)	113 (30.2)	
Skilled non-manual occupations	1213 (27.6)	110 (29.3)	
Skilled manual occupations	1218 (27.7)	104 (27.7)	
Partly-skilled occupations	424 (9.6)	43 (11.5)	
Unskilled occupations	128 (2.9)	5 (1.3)	
Maternal asthma	483 (10.6)	75 (19.4)	**<0.001**
Maternal eczema	1103 (24.3)	118 (30.5)	**0.008**
Maternal hayfever	1335 (29.4)	166 (42.9)	**<0.001**
Paternal asthma	343 (8.3)	64 (18.4)	**<0.001**
Paternal eczema	281 (6.9)	36 (10.4)	**0.019**
Older siblings	2061 (72.1)	171 (69.5)	0.425
No. of older siblings	1 (1.0)	1 (1.0)	0.359
Birth order			0.678
first-born	797 (27.9)	75 (30.5)	
second-born	1472 (51.5)	123 (50.0)	
≥3	589 (20.6)	48 (19.5)	
Maternal age at recruitment	29 (6.0)	29 (7.0)	**0.017**
Premature birth	222 (7.5)	24 (9.9)	0.215
Ever breast fed	3760 (81.9)	336 (86.2)	**0.040**
Vaginal delivery	4102 (89.2)	344 (87.8)	0.443
Age weaning (weeks)	13.8 (4.3)	13.5 (4.3)	0.418
Cat at 1 yr	1864 (42.5)	147 (39.3)	0.259
Cat at 4/5 yr	1809 (43.2)	136 (38.6)	0.106
Dog at 1 yr	1663 (37.9)	132 (35.3)	0.351
Dog at 4/5 yr	1590 (38.0)	127 (36.1)	0.513
Moved house by 2 yrs	705 (16.8)	68 (19.1)	0.292
Maternal smoking during pregnancy	525 (12.2)	61 (16.6)	**0.018**
Maternal smoking at 1 yr	1810 (39.4)	172 (44.3)	0.064
Current wheeze at 1 yr	1412 (31.0)	118 (32.2)	**<0.001**
Current wheeze at 2/3 yrs	465 (10.8)	123 (34.9)	**<0.001**
Current wheeze at 4/5 yrs	369 (8.9)	141 (40.1)	**<0.001**
Current wheeze at 8 or 10 yrs	274 (9.6)	41 (17.1)	**<0.001**
Frequency of wheezing at 1 yr			**<0.001**
Never	3355 (73.7)	229 (59.2)	
Sometimes	967 (21.2)	123 (31.8)	
Often	233 (5.1)	35 (9.2)	
Frequency of wheezing at 2/3 yrs			**<0.001**
Never	3682 (86.0)	240 (65.6)	
Sometimes	492 (11.5)	86 (23.5)	
Often	109 (2.5)	40 (10.9)	
Frequency of wheezing at 4/5 yrs			**<0.001**
Never	3803 (91.3)	212 (60.6)	
Sometimes	260 (6.2)	72 (20.6)	
Often	101 (2.4)	66 (18.9)	
Frequency of wheezing at 8 or 10 yrs			**<0.001**
Never	2585 (91.1)	199 (83.3)	
Sometimes	198 (7.0)	30 (12.6)	
Often	55 (1.9)	10 (4.2)	
Eczema at 1 yr	1201(27.2)	180 (47.9)	**<0.001**
Eczema at 2/3 yrs	1135 (26.5)	174 (47.2)	**<0.001**
Eczema at 4/5 yrs	1026 (24.5)	171 (47.4)	**<0.001**
DD asthma by 4/5 or 7.5 yrs	616 (15.0)	203 (60.8)	**<0.001**
Asthma treatment at 4/5 or 6.4 yrs	441 (10.8)	188 (54.0)	**<0.001**
SPT cat at 4/5 or 7 yrs	159 (4.7)	88 (29.4)	**<0.001**
SPT house dust mite at 4/5 or 7 yrs	338 (9.9)	114 (37.5)	**<0.001**
SPT mixed grass at 4/5 or 7 yrs	326 (9.6)	107 (35.5)	**<0.001**

Abbreviation - SES: socioeconomic status; yr(s): years; DD: doctor's diagnosis; SPT: skin prick testing

[a] *P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values ($p<0.05$).

Table D.4: Univariate analysis in the development data (ALSPAC and Ashford study) for the high-risk population

Candidate Predictor	Age 15–20		
	Non-asthma	Asthma	
	n (%) or median (IQR)	n (%) or median (IQR)	*P* value[a]
Sex (Female)	200 (50.6)	63 (45.3)	0.328
SES			0.750
Professional/managerial occupations	116 (30.6)	41 (30.4)	
Skilled non-manual occupations	96 (25.3)	40 (29.6)	
Skilled manual occupations	117 (30.9)	35 (25.9)	
Partly-skilled occupations	42 (11.1)	17 (12.6)	
Unskilled occupations	8 (2.1)	2 (1.5)	
Maternal asthma	66 (16.8)	36 (26.3)	**0.022**
Maternal eczema	130 (33.1)	48 (35.0)	0.754
Maternal hayfever	140 (35.6)	65 (47.4)	**0.019**
Paternal asthma	55 (15.1)	25 (19.4)	0.322
Paternal eczema	31 (8.6)	17 (13.2)	0.186
Older siblings	188 (69.6)	69 (71.1)	0.882
No. of older siblings	2 (2.0)	2 (1.0)	0.083
Birth order			0.886
first-born	82 (30.4)	28 (28.9)	
second-born	123 (45.6)	47 (48.4)	
≥3	65 (24.0)	22 (22.7)	
Maternal age at recruitment	29 (6.0)	30 (5.3)	0.510
Premature birth	25 (9.1)	7 (7.6)	0.824
Ever breast fed	314 (79.9)	118 (85.5)	0.184
Vaginal delivery	224 (82.1)	73 (81.1)	0.966
Age weaning (weeks)	13.04 (4.3)	13.04 (4.3)	0.244
Cat at 1 yr	166 (42.6)	45 (33.6)	0.084
Cat at 4/5 yrs	154 (40.0)	47 (35.1)	0.365
Dog at 1 yr	142 (36.4)	32 (23.9)	**0.011**
Dog at 4/5 yrs	148 (38.4)	41 (30.6)	0.128
Moved house by 2 yrs	72 (18.9)	24 (18.6)	>0.99
Maternal smoking during pregnancy	63 (17.1)	23 (17.0)	>0.99
Maternal smoking at 1 yr	157 (40.2)	64 (46.4)	0.240
Current wheeze at 1 yr	196 (49.6)	71 (51.1)	0.844
Current wheeze at 2/3 yrs	202 (51.1)	82 (59.0)	0.134
Current wheeze at 4/5 yrs	360 (89.8)	139 (97.9)	**0.004**
Current wheeze at 8 or 10 yrs	345 (87.6)	153 (98.2)	**<0.001**
Frequency of wheezing at 1 yr			0.449
Never	209 (52.9)	74 (53.6)	
Sometimes	132 (33.4)	40 (29.0)	
Often	54 (13.7)	24 (17.4)	
Frequency of wheezing at 2/3 yrs			0.208
Never	180 (48.1)	54 (40.7)	
Sometimes	141 (37.7)	53 (39.8)	
Often	53 (14.2)	26 (19.5)	
Frequency of wheezing at 4/5 yrs			**<0.001**
Never	41 (10.5)	3 (2.2)	
Sometimes	248 (63.8)	68 (50.4)	
Often	100 (25.7)	64 (47.4)	
Frequency of wheezing at 8 or 10 yrs			**<0.001**
Never	31 (8.3)	5 (4.2)	
Sometimes	201 (74.1)	48 (46.4)	
Often	85 (18.6)	56 (50.4)	
Eczema at 1 yr	153 (38.7)	80 (57.1)	**<0.001**
Eczema at 2/3 yrs	146 (38.6)	81 (60.0)	**<0.001**
Eczema at 4/5 yrs	155 (40.6)	87 (64.0)	**<0.001**
DD asthma by 4/5 or 7.5 yrs	188 (48.8)	18 (13.5)	**<0.001**
Asthma treatment at 4/5 or 6.4 yrs	235 (62.8)	113 (86.9)	**<0.001**
SPT cat at 4/5 or 7 yrs	51 (16.8)	45 (40.9)	**<0.001**
SPT house dust mite at 4/5 or 7 yrs	68 (22.4)	61 (53.5)	**<0.001**
SPT mixed grass at 4/5 or 7 yrs	63 (20.7)	53 (47.3)	**<0.001**

Abbreviation - SES: socioeconomic status; yr(s): years; DD: doctor's diagnosis; SPT: skin prick testing

[a] *P*-values obtained by chi-square test or Fisher's exact test. In bold, statistically significant *p*-values ($p<0.05$).