University of Southampton

Faculty of Medicine

Human Development and Health

Development of Childhood Asthma Prediction Models using Machine Learning and Data Integration

DOI: 10.5258/SOTON/D1943

by

Dilini M Kothalawala

ORCID ID 0000-0002-5804-0457

Thesis for the degree of Doctor of Philosophy

November 2021

University of Southampton

Abstract

Faculty of Medicine

Human Development and Health

Thesis for the degree of Doctor of Philosophy

Development of Childhood Asthma Prediction Models using Machine Learning and Data Integration

by

Dilini M Kothalawala

Childhood asthma is a chronic respiratory disease with substantial heterogeneity in its pathophysiology, presentation, trajectory and risk factors, particularly in early life. With the difficulty of obtaining an objective diagnosis before the age of five, the ability to predict childhood asthma could facilitate the identification of high-risk children, reduce misdiagnoses of probable asthmatics or encourage the implementation of primary prevention strategies and personalised asthma management. To promote the prediction of childhood asthma, a systematic review of existing prognostic prediction models for childhood asthma was conducted and demonstrated that current models have mainly been developed using traditional regression-based methods, with few independently validated and none being used in routine clinical practice. With the exploration of regression-based methods suggested to have been exhausted, this thesis aimed to explore novel approaches of data integration to improve current childhood asthma predictions using machine learning methods.

Using data from the Isle of Wight Birth Cohort (IOWBC, n=1456), the Childhood Asthma Prediction in Early-life (CAPE) and Childhood Asthma Prediction at Preschool-age (CAPP) models were developed to predict school-age asthma at 10 years using state-of-the-art machine learning methods. The CAPE and CAPP models used clinical and environmental data available from the first two year and first four years of life, respectively. Genome-wide genotype and methylation data were used to develop a polygenic risk score (PRS) and two novel methylation risk scores (MRS) (a newborn MRS, nMRS, and childhood MRS, cMRS) to predict childhood asthma, respectively. These genomic models were subsequently incorporated with the CAPE and CAPP models using a step-wise approach. The generalisability of all developed models was evaluated using data from the Manchester Asthma and Allergy Study (MAAS).

The CAPE and CAPP models demonstrated superior performance against their respective benchmark regression-based models based on area under the curve, with the CAPP model also surpassing the current best performing validated model, the Paediatric Asthma Risk Score (AUC: CAPE=0.71 vs. 0.64, CAPP=0.82 vs. PARS=0.80). The models offered good generalisability in MAAS and offered excellent sensitivity to predict a subgroup of individuals presenting with a persistent wheeze phenotype. Individually, the PRS and novel MRSs demonstrated moderate predictive ability (AUC: PRS=0.64, nMRS=0.61, cMRS=0.61). The integration of these genomic risk scores with the CAPE and CAPP models showed marginal improvement in performance (integrated CAPE=0.75, integrated CAPP=0.84). Overall, the incorporation of genetic and epigenetic data to predict the broad phenotype of asthma offered limited predictive improvement.

Using machine learning approaches, the CAPE and CAPP models were able to improve upon the current regression-based models for the prediction of childhood asthma. Coupled with the excellent sensitivity of the CAPE and CAPP models to predict a subgroup of individuals presenting with a persistent wheeze phenotype, this thesis suggests further exploration of the utility of machine learning methods focused on predicting asthma endotypes is warranted.

Table of Contents

Table of Tables

Table of Figures

List of Accompanying Materials

All material to support this thesis can be found at: https://doi.org/10.5258/SOTON/D1943

The accompanying material comprises of:

- Source code for all analyses conducted as part of this thesis.
- Datasets from the Isle of Wight Birth Cohort, namely the clinical data that was used to perform the analyses described in this thesis. Whilst supplied, access to this data is restricted and can be made available upon request from the David Hide Asthma & Allergy Research Centre. Further information can be found at www.allergyresearch.org.uk/.
- Supplementary results including: full descriptions of the candidate features considered during the development of the genomic risk scores for childhood asthma; full descriptions of the performance measures reported in the IOWBC for all candidate prediction models developed using machine learning approaches; all final trained childhood asthma prediction models to support the future application of the models developed in this thesis.
- Documentation for ethical approval, patient consent forms and participant information sheets from the IOWBC.

Research Thesis: Declaration of Authorship

Print name: Dilini M Kothalawala

Title of thesis: Development of Childhood Asthma Prediction Models using Machine Learning and Data Integration

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

Peer-reviewed articles:

Kothalawala DM, Kadalayil L, Weiss VBN, et al. Prediction models for childhood asthma: A systematic review. Pediatric Allergy and Immunology 2020;31:616-27. doi:10.1111/pai.13247.

Kothalawala DM, Kadalayil L, Weiss VBN, et al. Reply to Owora et al. Pediatric Allergy and Immunology 2020;32:393-5. doi:10.1111/pai.13396.

Kothalawala DM, Murray CS, Simpson A, et al. Development of childhood asthma prediction models using machine learning approaches. Clinical and Translational Allergy 2021;11. doi:10.1002/clt2.12076.

Preprint articles:

Kothalawala DM, Murray CS, Simpson A, et al. Development of childhood asthma prediction models using machine learning approaches. medRxiv 2021.03.31.21254678; doi:10.1101/2021.03.31.21254678.

Conference abstracts:

Kothalawala DM, Arshad SH, Holloway JW, et al. Abstract: Development of a childhood asthma prediction model using machine learning approaches. Allergy 2020;75:5-99. doi:10.1111/all.14504

Signature: .... Date: 11th November 2021

Acknowledgements

First and foremost, I would like to sincerely thank my supervisors Professor John W. Holloway, Dr Faisal I. Rezwan, Professor S. Hasan Arshad and Professor William J. Tapper for their expertise, guidance and support throughout this PhD. I am extremely grateful for their continued encouragement and willingness to share their knowledge; not only have they equipped me with the skills and confidence to successfully complete this PhD, but they have been valuable mentors for my personal and professional growth.

I would also like to give a special thanks to Dr Latha Kadalayil for her support and advice throughout this PhD. A sincere thanks to Dr Sadia Haider, Dr John Curtin, Dr Aref Kyyaly, Ms Veronique Weiss and Ms Paula Sands for their assistance and contributions to aspects of this research.

Additionally, I would like to acknowledge all the clinical and technical staff who have contributed towards the acquisition of data used in this thesis, particularly Nikki Graham, Stephen Potter, all the staff at the David Hide Asthma and Allergy Research Centre who undertook the assessments of the Isle of Wight birth cohort (IOWBC) as well as researchers of the STELAR/UNICORN Consortium who were involved with the data collection of the Manchester Asthma and Allergy Study (MAAS). A special thanks to all of the study participants and families of the IOWBC and MAAS birth cohorts, without whom this research would not be possible.

Most importantly, I would like to sincerely thank my family and friends, particularly my dearest parents, brother and sister, for their continued belief, patience and encouragement throughout this PhD – it has meant more than you know. Finally, I dedicate this thesis to my father, Mr N Kothalawala, whose sacrifices, unwavering strength and encouragement has driven my every success.

Definitions and Abbreviations

ADASYN	ADAptive SYNthetic sampling approach
API	Asthma Predictive Index
AUC	Area under the curve
BHR	Bronchial hyper-responsiveness
BMI	Body mass index
CAPE	Childhood Asthma Predictions in Early life
CAPP	Childhood Asthma Predictions at Preschool age
cMRS	Childhood methylation risk score
COPD	Chronic obstructive pulmonary disease
CpG	A site in the DNA sequence where a cytosine (C) is adjacent to a guanine (G) base, connected by a phosphodiester bond
DNA	Deoxyribonucleic acid
EWAS	Epigenome-wide association study
FeNO	Fractional exhaled nitric oxide
FEV_1	Forced expiratory volume in the first second
FN	False negative
FP	False positive
FPR	False positive rate
GINA	Global INitiative for Asthma
GWAS	Genome-wide association study
IBD	Identical by descent
ICS	Inhaled corticosteroids
IgE	Immunoglobulin E
INFO	Information metric used to quantify the quality of imputation of SNPs
IoW	Isle of Wight
IOWBC	Isle of Wight Birth Cohort

Definitions and Abbreviations

ISAAC	International Study of Asthma and Allergies in Childhood
KNN	K-nearest neighbours
LAASO	Least Absolute Shrinkage and Selection Operator
LR-	Negative likelihood ratio
LR+	Positive likelihood ratio
MAAS	Manchester Asthma and Allergy Study
MAF	Minor allele frequency
MICE	Multiple Imputation by Chain Equations
MRS	Methylation risk score
nMRS	Newborn methylation risk score
NPV	Negative predictive value
NRI	Net reclassification index
PARS	Paediatric Asthma Risk Score
PCA	Principle component analysis
PIAMA	Prevention and Incidence of Asthma and Mite Allergy
PPV	Positive predictive value
PRS	Polygenic risk score
RAST	Radioallergosorbent test
RBF	Radial basis function
RFE	Recursive Feature Elimination
ROC	Receiver operating characteristic curve
SABA	Short-acting beta agonist
SDS BMI	Body mass index standardised against the British Growth Reference
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SNP	Single nucleotide polymorphism
SPT	Skin prick test
SVM	Support vector machine

TN	True negative
TP	True positive
TPR	True positive rate

Chapter 1 Introduction

1.1 Childhood asthma

1.1.1 Defining asthma

Asthma is a chronic condition primarily affecting the conducting airways. However, asthma is not a single disease entity. Rather, it is a syndrome, used to describe a set of clinical characteristics, which may stem from a variety of pathological mechanisms[1-3]. Alongside variation in its pathophysiology, asthma presents with significant heterogeneity in its time of onset, presentation of symptoms, disease trajectory, severity, triggers and therapeutic response. As a result, establishing a single definition of asthma is challenging. The latest definition offered by the Global Initiative for Asthma (GINA) describes asthma as:

> "A heterogeneous disease, usually characterised by chronic airway inflammation. It is defined by the history of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable expiratory airflow limitation."
>
> **Global Initiative for Asthma (GINA) report, 2021[2]**

This definition offers only a general description based on typical clinical features of asthma that may distinguish it from other respiratory conditions. However, not only are these characteristics not observed in all asthmatics, they are also non-specific to asthma, resulting in a multitude of potential differential diagnoses[2,4]. Consequently, the presence of these characteristics merely act as criteria for deducing the probability of a patient's respiratory condition being asthma rather than confirming a diagnosis of asthma[5].

Despite sharing the same definition, childhood and adult asthma can be considered as two distinct forms of asthma[6,7]. For example, childhood asthma is associated with different phenotypes (observable characteristics) and risk factors, and characteristically presents with a male predominance before puberty, lower mortality and greater chance of remission compared to adult asthma[7]. Alongside the inherent heterogeneity of asthma, childhood asthma, the focus of this report, presents with an additional clinical complexity. This complexity may be accredited to limitations in diagnosing asthma in early life[2]. It may also be a result of the pathological mechanisms of paediatric asthma being entangled amongst the natural maturation of the respiratory and immune systems throughout childhood[2,5,8].

1.1.2 Burden of childhood asthma

Asthma affects approximately 339 million people worldwide, across all ages[2]. Whilst mortality remains low, asthma ranks 28th among the top global causes of disease burden when evaluated based on disability adjusted life years (DALYs), which accounts for both morbidity and mortality[9]. The distribution of the burden of asthma presents with bimodal peaks in childhood (age 10-14) and elderly age (age 75-89), although the prevalence of asthma is greatest in childhood[10]. In fact, asthma is the most common chronic disease in children, affecting 1 in 11 children in the UK (the main population focused on within this thesis)[11]. Notably, although the prevalence of childhood asthma is greater in high-income and westernised countries[3], severity is greater in low-income countries[12].

The economic burden of asthma can be categorised into direct and indirect costs. Direct costs are associated with the use of healthcare resources. In the UK, the National Health Service (NHS) spends approximately one billion pounds each year to care for asthmatic individuals of all ages[11]. Individuals with severe asthma, accounting for only 5% of the asthmatic population, consume 50% of the total asthma-related healthcare resources[13,14]. Specifically for preschool children with asthma or wheeze, a study conducted in 2003 estimated a total cost of 53 million pounds annually to the UK health service, 65% of which was expended on primary care[15]. Indirect costs refer to financial losses related to absence or reduced work productivity. A number of studies have reported that such indirect costs can exceed direct costs incurred (mainly in severe and uncontrolled asthma cases), and are what largely drive the economic burden of asthma, accounting for up to 75% of the total costs of asthma[16]. In the case of childhood asthma, indirect costs account for losses incurred by both the child and their parents/ caregivers.

In addition to the economic impact of the disease, childhood asthma can also have a large impact on the quality of life of both the patient and caregiver[17]. For example, emotional distress, absence from school/work, hospitalisation due to asthma exacerbations and limitations to normal activities can have a significant impact in reducing quality of life.

1.1.3 Pathophysiology

Asthma is characterised by airflow limitation caused by bronchoconstriction, airway hyper-responsiveness, oedema and chronic airway inflammation leading to a narrowing of the airway. Acute asthma is usually a result of bronchospasm -the spontaneous contraction of the bronchial wall muscle[1,2].

Airway hyper-responsiveness, a hallmark of asthma, refers to an exaggerated form of bronchoconstriction in response to various stimuli, structural changes to the airway and neuro-dysfunction[18-20]. Hyperplasia and hypersecretion of airway mucosal glands encourage the formation of obstructive mucosal plugs which can further obstruct and narrow the airway[18-20].

Inflammation in the airway can not only contribute towards the narrowing of the airway but can also encourage airway hypersensitivity, which further drives inflammation and hyper-responsiveness in the airway[18-22]. Airway obstruction observed in asthma is largely reversible, either spontaneously or supported by medication. However, chronic inflammation leading to airway remodelling can result in obstructions of the airway becoming (partly) irreversible[18]. For example, chronic inflammation can promote permanent structural changes through the thickening of the sub-basement membrane, fibrosis of the airway sub-epithelium, as well as hypertrophy and hyperplasia of airway smooth muscle[1,23].

However, these hallmark physiological changes observed in asthmatic patients can stem from a range of mechanistic pathways and be driven by a variety of genetic and environmental factors[19].

1.1.3.1 Asthma phenotypes and endotypes

The term asthma is limited by its non-specific description, accounting for a range of respiratory patterns. Childhood asthma usually manifests in infancy and persists into later life. However, symptoms can be persistent or intermittent, with severity improving or worsening with age and driven by different distinct underlying mechanisms (Figure 1.1)[2,24].

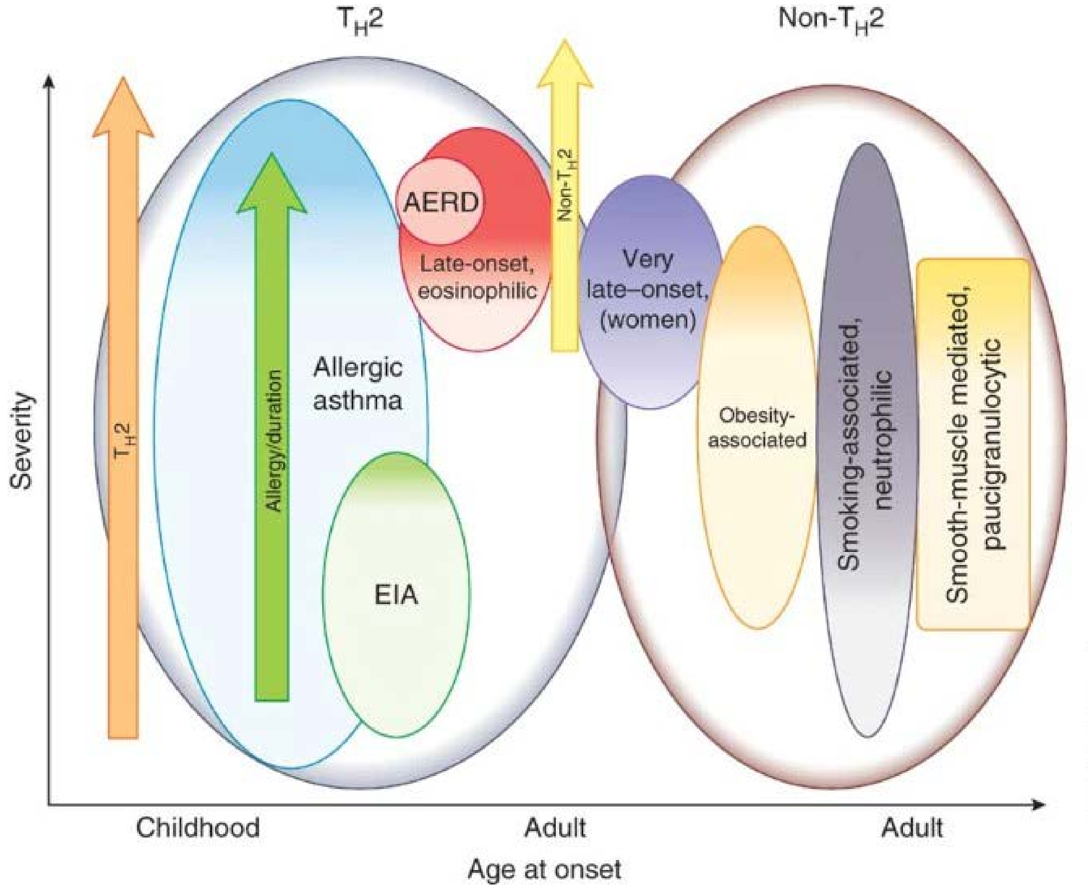


Figure 1.1 Asthma phenotypes and endotypes

Numerous phenotypes and endotypes of asthma have been proposed, characterised by variations in severity and age of onset. Figure adapted from Wenzel *et al.*[24], with permission from Springer Nature.

Phenotypes of asthma have been identified in order to untangle some of the heterogeneity observed in asthma. The initial classification of asthma identified two main phenotypes based on clinical observations for the time of onset and triggers of asthma – extrinsic (allergic) asthma and intrinsic (non-allergic) asthma[25]. Allergic asthma is characterised by early onset, high severity, a history of individual or familial allergic disease, atopy and identifiable triggers. Non-allergic asthma is characterised by late/adult onset, low severity, female-bias, and absence of allergic sensitisation[1,25]. In line with this early categorisation, childhood asthma has largely been considered an allergic disease of atopic pathology, and has established itself as the final feature of the atopic march (atopic dermatitis, allergic rhinitis and asthma) which is often present in atopic children[26].

Atopy refers to an exaggerated tendency to produce increased levels of immunoglobulin E (IgE) antibodies upon exposure to known allergens. Exposure to an allergen stimulates chronic eosinophilic airway inflammation and the production of specific IgE antibodies by B-lymphocytes. The allergen binds to specific IgE antibodies that are bound to mast cells and stimulates mast cell degranulation. The release of inflammatory mediators (e.g. histamine and leukotrienes) from mast cells in the airway epithelium stimulates airway inflammation and bronchial smooth muscle contraction[20,27].

However, studies suggest that only two-thirds of all asthma involves allergic mechanisms[28]. Furthermore, although atopic and non-atopic asthma can be identified as two clinical profiles, studies indicate similarities in their underlying pathologies[29]. For example, Humbert *et al*. identified similarities in the elevation of serum IgE, levels of T-helper (Th) 2 cytokines and response to inhaled corticosteroids (ICS) between allergic and non-allergic asthmatics[29]. Other suggested phenotypic groupings related to childhood asthma include late-onset eosinophilic, exercise-induced and neutrophilic asthma[24]. A number of unbiased clustering approaches have also been used to identify distinct phenotypes of childhood asthma[30-33] and wheeze[34,35]. Despite differences in terms of the methodologies and predictive features considered in each study resulting in a different number of clusters being identified, the identified clusters did display similar characteristics between studies.

The ability to stratify asthmatic patients based on their phenotypes in an unbiased way can offer some clinical insight, with studies attempting to identify differences in treatment response between clusters[33]. However, phenotypes do not provide insight into the underlying mechanisms of each stratum in order to guide targeted intervention. Rather, the classification of asthmatics into endotypes (defined as subtypes of a condition based on distinct functional pathophysiological mechanisms) may be of greater clinical use[24,36]. Two main endotypes have been identified in childhood asthmatics – Th2-high and Th2-low (Figure 1)[24]. The Th2-high subgroup characteristically presents with increased expression of Th2 pro-eosinophilic cytokines, atopy and sub-epithelial membrane thickening. In contrast, the Th2-low subgroup is characteristic of neutrophilic or paucigranulocytic inflammation. Unlike Th2-low, Th2-high asthmatics have been further identified as responsive to ICS. As a result, stratification of asthmatic patients into these endotype groupings could inform the implementation of therapeutic IgE or eosinophilic interventions. With the growing emergence of non-allergic asthma, a non-Th2 endotype has also been identified[24,25,37].

The relationship between asthma phenotypes and endotypes can be quite complex, reflective of the complexity of asthma itself. Each asthma phenotype may include more than one endotype; similarly, each endotype may account for multiple phenotypes[36,37]. However, whilst different asthma phenotypes and endotypes have been suggested, no consensus classifications have yet been established.

1.1.4 Diagnosis of childhood asthma

Due to the heterogeneous and non-specific expression of asthma, clinical guidelines are cautious in their diagnostic strategy, specifying an aim to deduce the probability for having asthma rather than to confirm a definitive diagnosis[5]. The British Thoracic Society and Scottish Intercollegiate Guidelines Network (BTS/SIGN) diagnostic algorithm for asthma comprises of a combination of clinical and objective tests[5]. However, both components present with high false positive and false negative rates for diagnosis.

Structured clinical tests comprise of a detailed family history of atopic disease, physical examination of symptoms and exposure to triggers. Clinical symptoms of asthma include wheezing, dyspnoea, chest tightness, coughing and nocturnal disturbances[5]. The experience of any of these symptoms in isolation generally offers poor predictive value, as the clinical symptoms of asthma are neither sensitive nor specific to asthma. For example, despite being the primary symptom observed in asthma, wheeze affects half of all preschool children. In these children, wheeze is often transient, with only one-third of individuals going on to develop asthma[38,39]. The probability of an asthma diagnosis increases with the presence of multiple symptoms. However, whilst most asthmatic children present with one or a combination of symptoms, only a quarter of children with asthma-like symptoms will have asthma whilst some individuals remain asymptomatic. For the remaining children displaying symptoms, alternative diagnoses include, but are not limited to, viral respiratory infection, cystic fibrosis, primary ciliary dyskinesia, developmental abnormalities or foreign body obstruction[5]. As a result, there is a substantial risk of misdiagnosis[3].

Lung function tests, including spirometry, fraction of exhaled nitric oxide (FeNO) in tidal breath and bronchoprovocation tests assess the hallmark physiological characteristics of asthma: variable airflow limitation, airway inflammation and bronchial hyper-responsiveness, respectively[2,5]. Despite being the most objective methods for diagnosing asthma, performing reproducible lung function tests on preschool children (<5 years old) is challenging[2].

Therefore, due to the difficulty of differentiating asthma from transient wheeze in early life, a clinical diagnosis of asthma cannot be reliably made until five years of age[2,8,18]. Prior to this, preschool children with a high or intermediate probability of asthma are initially identified based on an evaluation of a structured clinical assessment (presence of asthma-like symptoms, a family history of atopic disease and an absence of an alternative diagnosis)[5]. The diagnostic algorithm specifies two main approaches for diagnosing asthma in this group of suspected asthmatics – watchful waiting with review or monitored initiation of treatment. Due to viral respiratory infections being common at preschool age, children with mild, intermittent symptoms may be reviewed after a defined period, without intervention, to see whether the child's condition resolves by itself. In contrast, other symptomatic children may be offered a therapeutic trial of low-dose ICS and as-needed short-acting beta agonists (SABA) for 2-3 months. Children that demonstrate an improvement whilst on treatment, followed by worsening upon treatment cessation, are likely to be asthmatic[5].

1.1.5 Childhood asthma management

Despite the inability to provide a definitive diagnosis in childhood, preschool children presenting with asthma-like symptoms follow a similar management protocol as if diagnosed[3,40]. This comprises of both non-pharmacological and pharmacological interventions to achieve one of the three main goals of asthma management - primary prevention, secondary prevention or asthma control[5].

Asthma prevention is mainly driven through non-pharmacological interventions[5]. These include parent/patient education to reduce exposures to modifiable risk factors (detailed in Chapter 1.1.6) and addressing comorbidities of asthma. Non-pharmacological interventions can be implemented either before or after the onset of asthma, with the aim of reducing the incidence (primary prevention) or the impact of asthma (secondary prevention), respectively[5].

Conversely, pharmacological interventions aim to control asthma. Asthma guidelines describe a phased therapeutic approach, which targets treatment at a level corresponding to a patient's asthma severity in order to achieve early control[5]. The level of treatment can then be stepped-up or stepped-down to maintain control with the minimum pharmacological dose. Pharmacological intervention can be classified as reliever and controller medication. Reliever medication begins with SABAs for immediate relief of symptoms via bronchodilation. Controller medications aim to prevent symptoms from re-emerging. In preschool children, this may be achieved with the use of ICS or leukotriene receptor antagonists[2,5,14].

1.1.6 Risk factors for childhood asthma

Studies have identified a multitude of risk factors associated with the development of childhood asthma[41,42]. However, the heterogeneity of asthma also extends to its risk factors –different risk factors may contribute towards the risk of developing asthma or act as triggers for asthmatic episodes in different children. Risk factors of asthma include a genetic predisposition that cannot be changed (non-modifiable risk factor) as well as environmental exposures for which the size of the risk incurred may be altered (modifiable risk factor). One of the main goals of non-pharmacological interventions remains to alter the exposure to, and thus the subsequent effect of, modifiable risk factors in order to encourage primary and secondary childhood asthma prevention.

The contribution of genetic predisposition to the development of asthma has been widely established, with a family history of asthma and/or allergy being an important criteria used to identify preschool children at high risk of developing asthma in clinical practice[5]. Children with a history of familial asthma or allergy have been associated with an increased risk of developing asthma themselves[43-45]. The risk associated with parental asthma has shown to be additive, with the odds of developing childhood asthma being three and six times greater if one or both parents had asthma, respectively[46]. Maternal asthma and/or allergy has been found to confer a greater risk for the development of asthma compared to that observed from the paternal line[46]. However, Arshad *et al*. identified that the risk incurred by a parental history may in fact be a sex-linked association, with maternal asthma increasing the risk of childhood asthma in female children whilst paternal asthma conferred a greater risk in males. They also found a similar risk pattern with parental atopy[47]. Although the asthma risk associated with a family history of asthma or allergy primarily stems from a parental history of asthma, sibling associations have also been found (but these may also result from shared environment exposures)[48].

Reported estimates for the heritability of asthma from family and twin studies range between 25-80%[49]. Genome-Wide Association Studies (GWAS) have been extensively used to untangle the genetic predisposition of asthma, with 23 studies identifying associations of over 200 loci for childhood asthma onset[50]. The *GSDMB-ORMDL3* locus on chromosome 17q12-21 is the most replicated asthma locus to have been identified[51]. A recent GWAS conducted using data from UK

Biobank indicated that genetics may have a potentially more important role in the development of childhood asthma compared to adult asthma[52]. In this study, Pividori *et al.* identified a larger number of single nucleotide polymorphisms (SNPs), with overall larger effect sizes, to be associated with childhood onset asthma compared to adult onset asthma; of the 61 independent asthma loci identified, 23 were specific to childhood onset asthma and 37 loci were associated with both childhood and adult onset asthma[52]. Furthermore, based on the SNPs identified in the study, the estimated heritability was three times greater for childhood onset asthma (33%) than adult onset asthma (10%). Interestingly, genes related to adult onset asthma were highly expressed in the lungs and spleen whilst genes associated with childhood onset asthma were highly expressed in the skin. With the latter genes identified to be involved with epithelial barrier function, allergy and immune system regulation, it emphasises the association of childhood asthma with other allergic diseases such as atopic dermatitis[52].

Following the identification of these genetic biomarkers associated with childhood asthma, attempts to construct polygenic risk scores (PRS) to predict the risk of asthma have been made. Polygenic risk scores aim to estimate an individual's overall genetic risk through the summation of risk alleles (weighted by their allele effect size) across the genome. For example, Belsky *et al*. constructed a PRS consisting of 15 SNPs associated with asthma using data from the Dunedin Longitudinal Study birth cohort[53]. Based on their PRS, a one-standard-deviation increase in genetic risk corresponded to childhood onset (<age 9) asthmatic cases having a 20% higher risk of experiencing life-course persistent asthma (at age 13-38) (relative risk=1.20).

Epigenetic modifications, including DNA methylation, histone modifications, chromatin remodelling and non-coding RNA have been shown to alter gene activity which may subsequently contribute to the development of disease[54]. For example, Woodruff *et al*. identified the upregulated expression of the *POSTN*, *CLCA1* and *SERPINB2* genes to be associated with Th2 inflammation in samples of epithelial airway brushings in mild-to-moderate asthmatics compared to healthy controls[55]. Epigenetics may also explain some of the mechanisms underlying the complex gene-environmental interactions contributing to asthma development. Environmental exposures, such as traffic pollution and house dust mite, have been associated with differential DNA methylation of the Ten-Eleven Translocation 1 (*TET1)* enzyme, which has been suggested to be involved in the development of asthma[56]. Similarly, a range of other environmental risk factors, including, but not limited to breastfeeding[57], season of birth[58], urban living[59] and maternal smoking[60] have all been associated with altered methylation levels in genes related to both allergy and asthma.

Despite extensive research into the genetics of childhood asthma, many of the SNPs significantly associated with childhood asthma to date are common variants, present in both asthmatics and non-asthmatic individuals[61], and the proportion of estimated heritability accounted for by these common variants is lower than the total estimated heritability of childhood asthma[65]. This unaccounted heritability of disease is referred to as "missing heritability"[62,63]. A number of factors have been suggested to account for this, including the non-additive effects of genetic variants, complex gene-gene and gene-environment interactions, as well as genetic effects stemming from rare or yet to be discovered variants[62-64]. Suggestions that epigenetic modifications can be inherited have led to the epigenome also becoming a popular explanation for the missing heritability of risk present in childhood asthma[65]. Some studies have suggested that the inheritance of epigenetic signatures may persist across multiple generations[65], however, the extent of any such effects in humans remains unclear. As a result, a child's risk of developing asthma may be influenced by exposure to environmental risk factors, both directly by the child, but also indirectly through inherited epigenetic signals from parental or grandparental exposures. Such transgenerational effects may also contribute towards the missing heritability observed from GWAS.

The sex of an individual is also a potential risk factor for the development of asthma. In early life, males have an increased risk of developing asthma and present with more severe symptoms compared to females[66]. However, during puberty, a gender switch in the development of asthma is observed, with post-pubertal females demonstrating a greater risk of developing asthma, of greater severity, and requiring hospital admissions compared to males[67]. One suggested explanation for the observed gender reversal has been attributed to hormonal changes, whereby the increase in oestrogen exposure in females during puberty may drive airway inflammation[67].

Furthermore, childhood asthma is more prevalent in African American and Hispanic groups. Whilst some of the risk linked to ethnicity has been associated with genetic ancestry[49], individuals of these specific populations are likely to present with other risk factors of asthma such as low socioeconomic status, poor education and exposure to outdoor pollutants[42,68].

Tobacco smoke exposure is a well-established risk factor of asthma. Parental smoking, particularly maternal smoking either during or prior to pregnancy, significantly increases the risk of childhood asthma onset[69]. For example, maternal smoking during pregnancy was found to be associated with a 2.6 times increased risk of developing asthma in the first year of life[70]. Prenatal maternal smoking has further been associated with other risk factors of childhood asthma such as low birthweight, increased risk of recurrent chest infections, immune dysfunction and poor pulmonary development[71]. The effect of maternal smoking on childhood asthma has been suggested to be mediated through its role as an environmental pollutant, its impact on foetal growth as well as through epigenetic mechanisms[60,72].

In addition, allergy and atopy are well-described risk factors of childhood asthma development. Allergy describes the clinical manifestations following an exaggerated immune response to foreign antigens. Whilst different phenotypes of childhood asthma have been suggested, childhood asthma is still largely categorised as an allergic disease, particularly in early life[24]. Children presenting with other allergic diseases, primarily eczema or hay fever, have an increased risk of developing asthma. The atopic march describes the common pattern of development for this triad of allergic diseases observed in early life - first eczema, followed by hay fever and finally asthma[26], although studies in longitudinal birth cohorts have demonstrated that this sequential pattern is not typical for most children who develop asthma[73]. A parental history of any of these allergic diseases also increase the risk of childhood asthma development[44].

Conversely, atopy refers to the susceptibility to produce specific IgE antibodies in response to allergen exposure and has been identified as a hallmark characteristic of childhood onset asthma, used to distinguish between different asthma phenotypes. Sensitisation to environmental allergens is commonly detected through a skin prick test (SPT) or measure of specific IgE in blood through a radioallergosorbent test (RAST) or enzyme-linked immunosorbent assay (ELISA)[27]. Based on SPTs for a panel of 12 allergens, Arshad *et al*. identified that sensitisation to at least one allergen was significantly associated with a 4.56 times increased risk of developing asthma at preschool age (4 years old)[74]. The risk associated with allergic sensitisation demonstrated a linear relationship, whereby an individual's risk of developing asthma (as well as eczema and allergic rhinitis) increased with the number of positive SPTs. Although children were found to be sensitised to different allergens, the study identified that 94% of atopic children could be identified based on the four most common allergens (house dust mite, grass pollen, cat, and the fungus *Alternaria alternata*) alone. Sensitisation to the most common allergen, house dust mite, has further been associated with an eight times greater risk of developing asthma at 4 years[74]. A number of randomised controlled trials have shown that delayed exposure to house dust mite in early life is associated with a reduced risk of allergic sensitisation and allergic disease in childhood[41,75].

In contrast, recent randomised control trials, such as LEAP (Learning Early about Peanut Allergy)[76] and EAT (Enquiring about Tolerance)[77], have supported the recommendation for the early introduction of solid foods for the prevention of allergic disease, specifically food allergy. Between these trials, the early introduction of dietary foods resulted in a significantly lower prevalence of any food allergy, peanut allergy and egg allergy. In the EAT trial, early introduction of solid food at 12 and 36 months resulted in a 22% and 12% decreased risk of a child being atopic in the first three years of life, respectively (but these results were not statistically significant)[77]. Moreover, no significant associations were found in relation to the development of asthma, eczema or allergic rhinitis in these studies.

Despite conflicting findings reported in the literature, a number of systematic reviews and meta-analyses have identified significant protective benefits of breastfeeding for the development of asthma, early life wheeze and other allergic diseases[78-80]. Some studies suggest that these protective benefits are universal for all breastfed children whilst others indicate only a skewed benefit towards infants with a genetic predisposition for allergic diseases[78-80]. These protective benefits may stem directly from breastmilk exposure through its known immunoregulatory effects, or indirectly by promoting cow's milk protein allergen avoidance, or both[75,81]. Furthermore, the complex composition of breastmilk compared to formula food may promote immune tolerance by facilitating the colonisation of a more diverse gut microbiota[82].

The hygiene hypothesis suggests that reduced microbial exposure in early life can hinder the development of the immune system, reducing a child's immune tolerance and increasing their susceptibility for allergic and other immunological diseases[83,84]. Improved sanitation, dietary changes and widespread immunisation are factors suggested to reduce microbial exposure. As a result, the hygiene hypothesis has been used as a suggested explanation for the increasing prevalence of asthma observed in westernised countries[83,84]. Factors such as: breastfeeding[82], increased family size or number of siblings[48], vaginal delivery[85], living on a farm in early life[86], exposure to household pets and day-care attendance[48], all of which promote microbial exposure within the first few years of life, have been associated as protective risk factors of childhood asthma[42,83].

Whilst early life microbial exposures may offer protective benefits against the development of childhood asthma, lower respiratory tract infections with respiratory syncytial virus (RSV) and human rhinovirus (HRV) have been significantly associated with persistent wheeze within the first five years of life and are common causes of hospitalisation[87]. Jackson *et al*. identified that children who experienced wheeze due to RSV or HRV infections in the first three years of life were approximately 3 and 10 times more likely to develop asthma at age 6, respectively[88].

Other risk factors related to foetal growth, such as childhood obesity, birthweight, prematurity and maternal age, have also been suggested as potential risk factors of childhood asthma[41,69]. Based on studies that have identified distinct body mass index (BMI) trajectories in childhood, a rapid increase in BMI within the first two years of life as well as those trajectories indicative of childhood obesity, were associated with an increased risk of childhood asthma development[89,90]. Despite conflicting results, some studies have identified low birthweight, prematurity and maternal age to be significantly associated with childhood asthma, possibly due to their impact on infant lung growth and susceptibility to respiratory infections[91,92].

Indeed, clinical symptoms are the most obvious risk factors for diagnosing asthma. Clinical symptoms, such as recurrent wheeze, cough and nocturnal symptoms, observed in early life have shown to be predictive risk factors for the future development of asthma at school age, with the presence of multiple symptoms often conferring an increased risk[5]. Therefore, the consideration of clinical symptoms alongside both modifiable and non-modifiable risk factors is important to assist in the diagnosis, prediction and potential prevention of childhood asthma.

1.2 Predicting health outcomes

With the economic burden of many chronic non-communicable diseases, such as asthma, rising, already challenged healthcare services need to ensure the efficient and cost-effective utilisation of healthcare resources whilst continuing to improve the accuracy of diagnosis and treatment[93]. Predictive risk modelling of diseases has become a prominent area of research in healthcare – utilised for the purpose of forecasting widespread, highly burdensome diseases, such as seasonal influenza[94] or global pandemics[95,96], diagnosing diseases and generating individual risk predictions[97].

Prediction models can either be diagnostic (estimating the probability of currently having the outcome) or prognostic (estimating the probability of developing the outcome in the future)[98]. For both, the aim of individual predictive risk modelling in healthcare is to provide an objective measure to support physicians in their clinical decision-making. Identifying individuals at risk of developing disease in an accurate and more objective manner may enable physicians to initiate early prevention strategies and efficiently direct personalised, targeted care towards those most at risk. As a result, predictive models could help to reduce unnecessary wastage of healthcare resources as well as limit unnecessary exposure to treatments and their associated risk of adverse effects. Predictive modelling could further encourage individual awareness of one's own risk of disease. Engaging individuals with a quantifiable risk prediction may motivate significant lifestyle changes, avoidance of potential modifiable risk factors and improve adherence to treatments. Such patient engagement could help individuals to curb their natural disease trajectory more effectively and manage their disease outside the bounds of healthcare services[93,99-101].

1.2.1 Predictive risk models in other disease areas

The development of predictive risk models has rapidly increased across a variety of healthcare areas. A systematic review conducted in 2011 identified that new predictive risk models were being developed at a rate of one every three weeks[99]. Models for predicting an individual's risk of disease were first developed for cardiovascular disease (CVD); to date, over 350 different prediction models have been developed[102]. For example, the Framingham Score is one of the most established risk scores for CVD in clinical practice and is suggested in numerous guidelines as an assistive tool for clinical decision-making[103]. In addition, predictive risk models have been developed in adults for type 2 diabetes[99], hospital readmissions[104], acute chronic obstructive pulmonary disease (COPD) exacerbations[105,106] and post-transplantation outcomes[107], to name a few. Predictive modelling studies specifically focusing on paediatric patients have also been developed in areas such as oncology[108], respiratory conditions[109] and childhood obesity[110]. However, in some disciplines, such as obesity, only a few paediatric models have been implemented into clinical practice for predicting the risk of childhood disease[110].

1.2.2 The need for predicting childhood asthma

Approximately 80% of childhood asthmatics develop symptoms before the age of six[7,111]. However, there are no objective respiratory tests available to accurately diagnose asthma before this age. As a result, children who present with asthma-like symptoms in these early years of life are either diagnosed as having viral induced wheeze and left untreated or considered as probable asthmatics and prescribed asthma medications.

Some studies suggest that there is a current transition from an era of under-diagnosis - where asthmatic children were often untreated and at risk of developing more severe asthma, to an era of over-diagnosis – where increasingly, non-asthmatic children are cautiously treated as probable asthmatics, resulting in unnecessary exposure to treatment and increased burden on healthcare services[112,113]. For example, despite their effective management of asthma-symptoms, treatment with ICS has been associated with impaired bone development and stunted growth velocity in children within the first two years of use, even at low doses[114]. Yet, both under and over-diagnoses of childhood asthma are common occurrences.

Prediction models for asthma can help to identify preschool children at high-risk of developing asthma at school-age and distinguish them from those whose symptoms are transient. Early knowledge of a child's asthma risk may offer physicians support in deciding to prescribe or withhold medication[112]. Risk prediction tools can also be powerful mechanisms for promoting asthma management outside of healthcare service intervention by allowing parents to understand their child's asthma risk in a quantifiable and meaningful way. Additionally, prediction tools can enable physicians to initiate a conversation with parents and encourage methods of actively managing their child's pulmonary health, e.g. by reducing their child's exposure to modifiable risk factors[112,115]. Childhood asthma and poor lung function in early childhood have also been strongly associated with the development of COPD in later life[116]. Therefore, informed actions taken following the identification of children at risk of developing childhood asthma may not only help to prevent or curb the progression of the child's asthma throughout childhood, but may also reduce the risk of developing other respiratory diseases later in adulthood.

1.2.3 Methods to predict health outcomes

The aim of predictive modelling is to provide insight into the probability of an event outcome rather than to determine causality between features (predictors) and an outcome[117]. The process of predictive modelling comprises of development and application. The development process of a prediction model is referred to as the training phase, in which a model is chosen and optimised to predict an outcome of choice based on a representative dataset[118]. The optimised model is then applied to make predictions on new data. Although there is a large variety of methods which can be employed for predictive modelling, they can be broadly classified into two approaches - statistical and machine learning approaches.

The primary aim of traditional statistical methods is to make inferences on the relationship between variables and the outcome using existing insight of the problem domain and data from a representative sample of the problem population. These are generally parametric methods, where assumptions are made on the underlying distribution of the data. Traditional statistical methods can then be used to make predictions based on these inferences[118].

Regression models are the most commonly applied statistical approach for predictive modelling. The simplest regression method is the simple linear regression, whereby predictions of a dependent variable (outcome) are made based on a linear relationship between itself and a single independent variable (predictor). In a regression analysis, the relationship between the predictor and outcome is quantified by the regression coefficient. In many situations, including healthcare, multiple variables are related to an outcome. Hence, multiple linear regression, which predicts an outcome based on its relationship with a set of predictors, is more commonly used. In this situation, the regression coefficient for each predictor is calculated based on the average effect of a unit increase of the predictor on the outcome, fixing the effect of all the other predictors included in the model. It is assumed that each predictor included in the model has an additive influence on the outcome. However, it is possible to incorporate interaction terms into the model to relax this additive assumption. Linear regression is used when the outcome is a continuous variable whilst logistic regression is used for binary outcomes[118].

However, there are important limitations of regression models. First, regression methods assume that the data is linearly separable. If the data is not truly linearly separable, the regression model will demonstrate poor predictive accuracy. Another limitation centres on the assumption that error terms for each observation are uncorrelated and have constant variance across the dependent variable. Deviations from these assumptions can result in predictors erroneously being considered statistically significant and an underestimation in the true standard error of the model. Regression models are also highly affected by outliers and extreme values. Finally, an important limitation of regression models is that they are also affected by multicollinearity between predictors which are highly correlated with each other. The collinearity between predictors can reduce the ability of a model to identify significant predictors and can lead to erroneous estimates of the individual effect size of each predictor on the outcome[118].

In contrast to traditional statistical methods, the primary goal of machine learning approaches is to make accurate predictions. This is achieved by machine learning algorithms recognising relationships present within a subset of training data and evaluating the predictive performance of the model on a separate test dataset. The focus is on whether relationships are present within the data, with less concern over understanding these relationships. It is this distinction which has driven the exploration of machine learning across a variety of fields – compared to statistical methods, they are better able to address the complex, often non-linear relationships that underpin many real-world problems[117,118].

A variety of machine learning algorithms exist, including both parametric and non-parametric (no assumptions made on the underlying distribution of the data) algorithms. These algorithms range from highly interpretable models with low complexity to highly complex, black-box models which are difficult to interpret[119]. However, due to the focus on making accurate predictions rather than explaining identified relationships between predictors and the outcome, many machine learning approaches often sacrifice model interpretability for an improvement in prediction accuracy. It is important to note that there is no consensus distinction between traditional regression-based and machine learning-based models. Whilst some distinguish these methods by whether the prediction outcome is continuous or classifying distinct classes, others consider whether the methods are parametric or non-parametric. In addition, many do not consider simpler models such as logistic regression as machine learning models. Throughout this thesis, traditional statistical and regression-based models refer to those models which require assumptions to be made on the distribution of the data, including logistic regression models. Machine learning-based models refer to those which use non-parametric and more complex algorithms.

Particularly with the growth of electronic health records, access to big data has fuelled the popularity of utilising machine learning methodologies in healthcare[120]. Given the complex interactions between biological variables, the exploration of machine learning methods which focus merely on identifying relationships within data rather than attempting to untangle mechanisms of complex interactions, may hold promise in improving prediction accuracy compared to traditional statistical methods. Numerous studies have compared the utilisation of these two approaches of predictive modelling for healthcare; although not always consistent, machine learning approaches have demonstrated comparable, if not superior, predictive performance compared to traditional statistical methods across a number of disease areas[121-123].

1.3 Machine learning

Machine learning is a branch of artificial intelligence that utilises principles of statistics, mathematics and computer science to develop algorithms that learn directly from data and experience[118]. Algorithms under the umbrella of machine learning can be broadly classified into two categories: supervised and unsupervised learning. Supervised learning methods utilise labelled data, whereby each observation has data for both a set of predictive features and an assignment of its outcome. Supervised learning methods can be further classified based on the type of outcome being predicted – classification algorithms predicting outcomes into discrete categories and regression algorithms predicting continuous outcomes. Conversely, unsupervised learning methods aim to identify underlying patterns within data, without any prior information about the outcome. Examples of unsupervised learning include clustering and dimensionality reduction. There is also another branch of machine learning, called semi-supervised learning. This refers to a process of supervised learning conducted on a small initial training subset, followed by unsupervised learning on the remaining training dataset[118,124].

1.3.1 Predictive modelling for classification

Prediction models developed for classification purposes using machine learning are an example of supervised machine learning. The process comprises of two phases: training – to develop and optimise a learning algorithm; and testing – to evaluate the performance and generalisability of the model on unseen data. When developing a classification model, the model learns patterns from the training data and is evaluated on the test set[118,119].

Machine learning algorithms can be described based on two main criteria – interpretability and flexibility (Figure 1.2)[118]. Interpretability refers to the ability to understand how the interactions between predictors led to the classification output. Model flexibility refers to the ability for an algorithm to fit numerous forms of its defining function, optimising itself based on the complexity of the data. For example, when linear regression is applied for the purpose of making predictions rather than inferences, it may be considered to be a machine learning method[118]. Models developed using linear regression are considered to be easily interpretable due to the additive effect of each predictor, quantified by the regression coefficient. But these models are inflexible - they are limited to construct linear decision boundaries, irrespective of whether the data is in fact linearly separable. In contrast, support vector machines (SVMs) are considered highly flexible models, with the potential to apply different kernel functions to accommodate complex data patterns. Kernel functions allow complex, non-linearly separable data to be mapped into higher dimensional spaces where they become separable[118]. However, this complexity makes it difficult to understand how the output of the SVM was derived from the input variables provided. Hence, SVMs are considered to be difficult to interpret "black-box" algorithms.

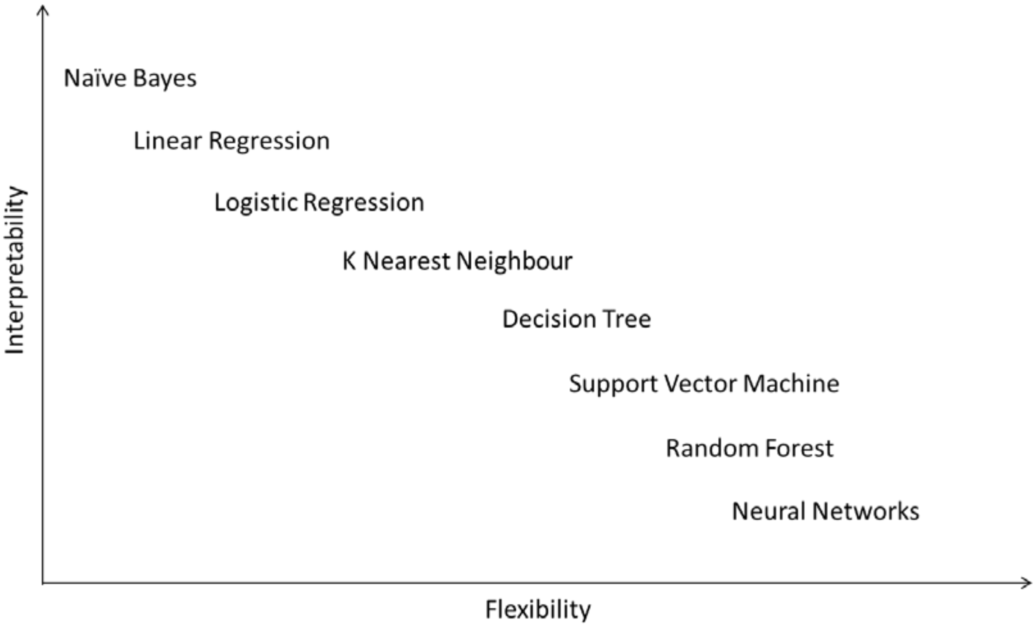


Figure 1.2 The interpretability-flexibility trade-off for supervised machine learning algorithms

Algorithms that offer a greater degree of flexibility tend to have lower model interpretability. Figure adapted from James *et al*.[118], with permission from Springer Nature.

Irrespective of the algorithm used, the training process aims to, either directly or indirectly, reduce the error in the training data[118]. As a result, it is expected that performance in the training dataset will be superior to that of the test set. This concept can be exploited to identify problems of overfitting in a developed model. Overfitting refers to the process of a model learning the patterns of the training data too well such that, when applied to a set of similar unseen data, the model is unable to make accurate predictions[125]. In practice, this would be evident from a low classification error in the training data but a large error in the test data. Models subject to overfitting are not generalisable.

The problem of overfitting stems from the trade-off between bias and variance during model development (Figure 1.3). Bias refers to the ability of a model to make accurate real-life approximations. In contrast, variance refers to the extent to which a model would change given alterations to the training data[118]. Ideally, a model should have low bias and low variance. However, there is always a trade-off between these two parameters. For example, in the case of a non-linear classification problem, an inflexible model, such as one developed by linear regression, may not be able to account for the full complexity of the data with a simple linear separation. Such underfitting of the training data can result in a model with high bias[118]. However, this model may present with low variance due to small changes in the training data only having a small influence on the position of the decision boundary and the subsequent classifications made. In contrast, models developed by SVMs for example, have greater flexibility to recognise and learn complex patterns in the data, thus improving the classification accuracy and reducing model bias. However, in order to accommodate the complexity of the data, the model becomes dependent on the properties of the training dataset. As a result, the model will have high variance, with small changes in the training data potentially having a large impact on the decision boundary[118].

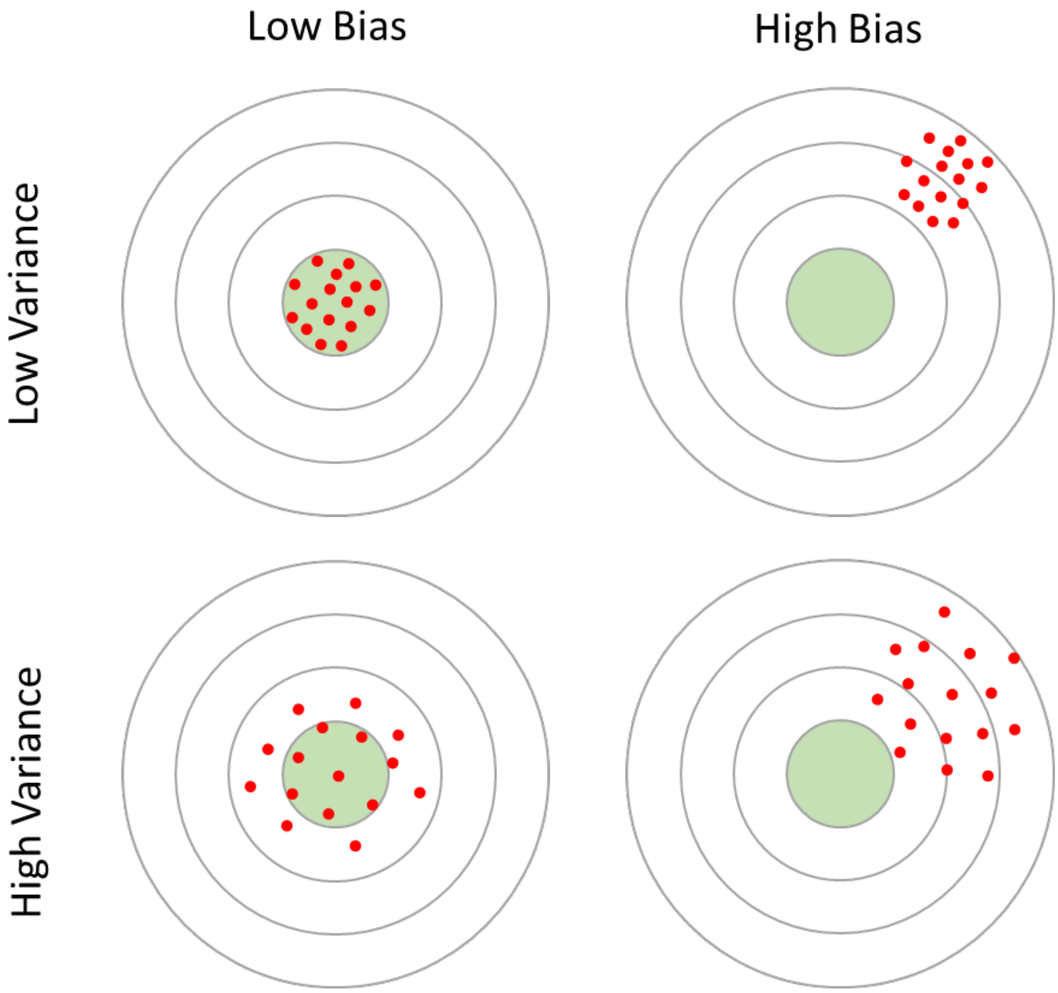


Figure 1.3 The bias-variance trade-off for classification model predictions

In each target space, the green circle represents accurate predictions. Each target depicts one of the four main prediction patterns (red dots) that can be offered by a classification model in terms of bias and variance. Whilst the top-left target shows the ideal prediction pattern of low bias and low variance, prediction models generally present as one of other targets due to the bias-variance trade off. Figure reproduced based on Doroudi *et al*.[126] (copyright license: CC-BY-NC 4.0).

A number of approaches to address the common problem of overfitting and establish a balance between model bias and variance have been suggested[118]. These approaches include using datasets with large sample sizes (>1000 samples) where available; using resampling methods (e.g. cross validation or bootstrapping) where available datasets are small; reducing noise in the data by applying feature selection methods; and encouraging unbiased training using techniques for class imbalance.

1.3.2 Resampling methods

The bias-variance trade-off is a significant problem, particularly when models are developed using small datasets. Withholding a portion of the dataset for testing purposes further reduces the number of observations available for model training. To address this, resampling techniques, such as cross validation and bootstrapping, are used to artificially increase the number of observations used for model training from the available data[118]. Cross-validation is an iterative process whereby all observations are used for both training and testing. In this process, the data is divided into *k*-folds; a model is trained on (*k-1*) folds and evaluated on the remaining fold. The training and testing process is repeated *k*-times using a different fold as the test set each time (Figure 1.4). This process can be applied to identify tuning parameters in a bid to construct a generalisable model and reduce overfitting. This process can also be used to establish a generalised estimate of the performance of a model.

Similarly, bootstrapping is the process whereby multiple training datasets of equal size are produced from the original dataset. This technique resamples with replacement - as a result, approximately 80% of the original dataset forms a new dataset for model training. The remaining 20%, known as the out-of-bag samples, are used as the test set to evaluate the performance of a model[118]. With both resampling techniques, the number of observations used to train a model is maximised in order to better recognise patterns and make accurate predictions (reducing bias). In addition, by training a model on multiple sets of data, the model becomes less dependent on the properties of the training data, lowering the potential for overfitting (reducing variance) and increasing the generalisability of the model.

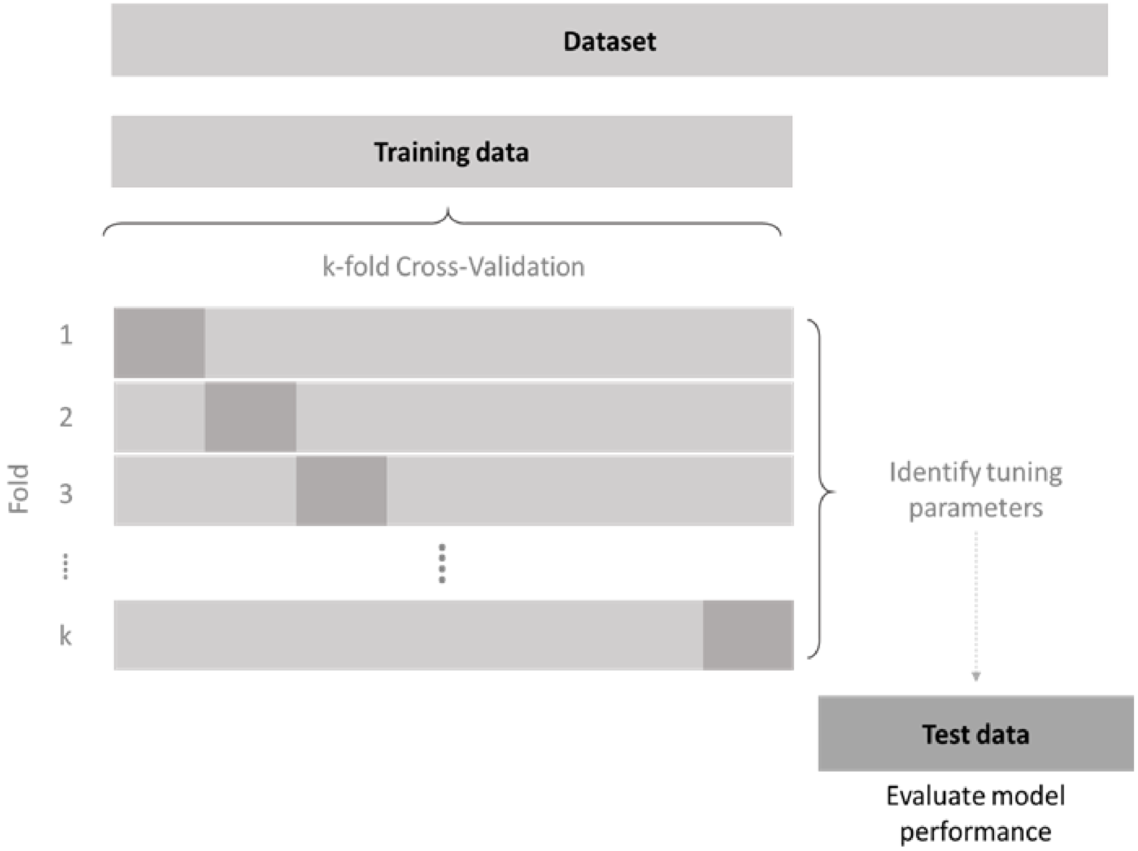


Figure 1.4 Example of how cross-validation may be used to develop a machine learning model

The dataset is split into a training and test set. Using the training set, all aspects related to model development (such as the tuning of model parameters) are performed within a cross-validation framework. Based on the collective results obtained from each cross-validation fold, the final model is defined and applied to the unseen test data. An evaluation of the model on the test data provides an indication of how well the model performs and generalises. Figure reproduced from Pedregosa et al.[127] (copyright license: BSD).

1.3.3 Feature selection

A fundamental consideration for modelling is the input data that is used. Technological advances and data storage capabilities have facilitated the collection of large amounts of data with increasing ease and speed[118]. As a result, researchers are increasingly faced with large amounts of potential input data available for analyses. Additionally, data is often heterogeneous, comprising of different data types. Whilst the combination of different data types can offer a holistic understanding of the problem in question compared to single-data-type analyses (particularly in areas such as healthcare and bioinformatics), data integration has often posed a significant hurdle[128]. Whilst advancements in machine learning have facilitated a number of approaches to tackle data integration, there are still many challenges to integrate heterogeneous data types (e.g. the curse of dimensionality), with no gold-standard method identified[128,129].

Regardless of whether data is heterogeneous or of a single data type, the utilisation of high dimensional data is a known problem in statistical modelling. High dimensional data refers to an input feature space where the number of features is very large, often much larger than the number of observations[118]. Not only is this computationally costly, it can also subject models to the "curse of dimensionality". This refers to the phenomenon whereby, as the number of features considered in a model increases, noise introduced into the training data encourages the model to overfit on the training data. This subsequently compromises the test performance and future generalisability of the model[130].

Despite the availability of data for a large number of features, not all available features are required for modelling a specific problem. Candidate features can be evaluated for their relevance and redundancy. Feature relevance refers to the strength of the relationship between the feature and the outcome. In contrast, redundancy evaluates the relationship between features; a redundant feature is one that is dependent or often highly correlated with another feature that can fully explain its relationship with the outcome[131]. Based on these evaluation criteria, methods have been proposed to reduce the dimensionality of data during model development. These methods can be classified into: feature selection – which identifies a subset of the most informative candidate features from the original feature set; and feature extraction – which performs transformations on the original features to extract a reduced set of new features that can explain the patterns of the original feature set[131].

Feature selection methods aim to make the distinction between relevant and redundant features in order to obtain a subset of useful features to be carried forward for modelling. These methods are often more interpretable than feature extraction methods. There are three main types of feature selection methods – filter, wrapper and embedded methods[131]. Filter methods, such as mutual information-based feature selection, information gain or relief, have the lowest computational cost and are independent of machine learning algorithms[132]. These methods evaluate feature relevance based on a specified evaluation criterion, such as Akaike information criterion, Bayesian information criterion, mean squared error or correlation coefficients[131]. Wrapper methods, such as Recursive Feature Elimination (RFE) or genetic algorithms, are computationally expensive feature selection methods; they train a chosen modelling algorithm on each possible subset of features to identify the subset of features which offer the maximal predictive accuracy[133]. Finally, embedded methods perform feature selection within the model construction process and are usually model specific. For example, least absolute shrinkage and selection operator (LASSO) and Elastic Net employ a regularisation step within a regression model in order to shrink the coefficients of less important features[131]. In this way, embedded methods are able to reduce dimensionality whilst retaining all candidate features as inputs in the model. Compared to traditional statistical methods, machine learning approaches are more robust to concerns of multicollinearity between predictors. As a result, machine learning approaches utilising robust feature selection methods may offer the opportunity to recognise patterns within data and potentially identify novel predictors, which may have been previously overlooked by regression-based approaches[117,118,123].

1.3.4 Addressing the data imbalance problem

The fundamental aspect of supervised machine learning is the process of learning from given data examples to predict unseen data[118,119]. In the case of a two-class problem, optimal learning would occur when an equal number of examples of both classes were available for the algorithm to learn. In fact, most algorithms assume that the distribution of classes are balanced and the cost of misclassification is consistent between classes[134]. However, obtaining a perfectly balanced dataset is unrealistic in many real-world settings. Often in classification problems, particularly in healthcare, the event of interest belongs to the minority class. For data with a significant class imbalance, the accuracy of modelling algorithms can be substantially limited due to the bias towards predicting the majority class[134].

The problem of imbalanced data is well documented, with numerous methods suggested to address the problem[134]. Whilst the obvious solution would be to collect additional data for examples of the minority class, this is not always feasible depending on research costs (extrinsic imbalances) or the nature of the dataspace e.g. low event prevalence (intrinsic imbalances)[134]. Other approaches addressing this issue include sampling methods, the generation of synthetic data, cost-sensitive learning, active learning and kernel-based methods[134]. The utilisation of rebalancing methods has been shown to improve the performance of modelling algorithms and should not be overlooked when dealing with imbalanced data[135].

Sampling methods involve introducing or removing examples of the original dataset in a random or informed manner. Oversampling refers to the replication of examples belonging to the minority class, which are then added to the original training dataset. Undersampling involves the removal of examples of the majority class from the original training dataset. Whilst undersampling can improve the class imbalance, this method forces a decrease in sample size and the removal of potentially useful training data, hence is not often preferred[134].

Although replicating existing examples of the minority class addresses the data imbalance, the addition of duplicate examples can result in overfitting[134]. To overcome this problem, oversampling methods, which generate synthetic data for the minority class, have been proposed. The Synthetic Minority Oversampling TEchnique (SMOTE), which generates synthetic examples based on the *k*-nearest neighbours of existing minority examples, has shown to be highly effective and is widely implemented[136]. Suggested improvements upon SMOTE include borderline-SMOTE[137] and ADAptive SYNthetic sampling (ADASYN)[138]. These methods also generate synthetic examples but focus on increasing the number of difficult to classify minority examples. Other methods to address the data imbalance problem, which do not require alterations to the dataset, have been proposed. One example is cost-sensitive learning, whereby the algorithm penalises classifications based on a cost matrix. Often, penalties are only incurred on misclassifications, with higher penalties given for misclassifications of the minority class compared to the majority class[139].

Another important consideration for handling imbalanced data is the correct use and reporting of performance measures[134]. The success of modelling algorithms is generally evaluated by a measure of accuracy, the proportion of correct classifications. However, accuracy can be misleading when dealing with imbalanced data. For example, consider a two-class classification problem in which all examples of the minority class were misclassified. For a balanced dataset, the reported accuracy would be 50%, clearly highlighting that the performance of the model was no better than chance. However, for an imbalanced dataset with only one-tenth of examples belonging to the minority class, the reported accuracy would be 90%, suggesting at first glance that the model offered excellent predictions. In such situations, reporting the balanced accuracy (the average proportion of correct classifications for each class) may be more appropriate. Other performance measures, such as the F_1-score, recall, precision and area under the receiver operating characteristic (ROC) curve are less influenced by class imbalances and should also be reported in such situations (discussed in Chapter 2.3.6)[134].

1.3.5 Handling missing data

Missing data is a common problem that can influence both the development and application of modelling algorithms. There are three main patterns of missing data. First, data can be "missing completely at random" (MCAR), where there is no underlying systematic difference between missing and non-missing data. Second, data can be "missing at random" (MAR), where any underlying systematic differences between missing and non-missing data can be explained by differences in the non-missing data. Finally, data can be "missing not at random" (MNAR), where any systematic differences identified between the missing and non-missing data cannot be explained by the non-missing data[140-142].

A simple solution to deal with missing data is to remove observations with missing data and proceed with complete data analyses. However, such an approach can introduce bias into the analyses, particularly if data is MNAR[140]. Complete data analyses can also substantially reduce sample size and study power, resulting in a potential loss of precision and under-fitting[142].

There are numerous methods for handling missing data[140,142-144]. Simple approaches include single imputation methods such as mean imputation or last measure carried forward[144]. More complex approaches include multiple imputation methods such as Multivariate Imputation by Chain Equation (MICE), which involves the generation of multiple imputed datasets upon which the results of an analysis are pooled across each dataset[145,146]. Such methods aim to address the inherent uncertainty of the missing data values. Other methods include maximal likelihood estimation and approaches utilising machine learning algorithms, such as nearest neighbour estimation[144,147] and missForest[148].

Despite the potential benefits of imputing missing data, imputation should be performed with caution when there is a substantial proportion of missing data[141]. It has been suggested that analyses may be subject to bias when the proportion of missing data exceeds 10%[141,142]. However, a few studies have showed that very large proportions of missing data alone may not introduce bias, emphasising that other considerations also need to be made, such as assumptions of missing data patterns[141] or the fraction of missing information[149].

1.3.6 Use of machine learning for disease prediction

Machine learning has been widely applied across healthcare to address medical problems including: disease diagnosis[122,150], predicting health outcomes, such as mortality, the development of disease or other comorbidities[151-153], as well as predicting adherence to treatment[154] and the utilisation of healthcare resources[155]. Ensemble machine learning approaches (e.g. random forest) have commonly shown robust performance as classification tools in disease areas, such as Alzheimer's disease[156] and cardiovascular disease[153]. SVM, a more recently developed machine learning approach, has also established itself as a powerful classification tool and has been applied to the prediction of diseases such as diabetes[151] and inflammatory bowel disease[157]. In addition, unsupervised clustering approaches have been applied to stratify patients with similarly presenting conditions[157] as well as to identify different disease and comorbidity trajectories[152].

Particularly within healthcare, data is often heterogeneous, sourced from different modalities, such as questionnaires, images, recordings and a variety of omics analyses. Although methods that are able to integrate multiple data structures for single analyses are not yet well established, machine learning approaches used to integrate heterogeneous data have been applied to a number of areas of biomedicine[129].

1.3.7 Use of machine learning in asthma

Supervised machine learning approaches have also been applied to predict the presence of asthma[158], as well as the chance of future asthma exacerbations[159]. Studies have also applied machine learning methods to optimise the management of healthcare resources, for example by predicting post-exacerbation hospitalisation and clinical decision making at triage in adults presenting to the emergency department with asthma exacerbations[105].

Unsupervised machine learning approaches have also been employed in an attempt to untangle some of the underlying heterogeneity seen in the pathophysiology and treatment of asthma; by stratifying asthmatic individuals, personalised asthma care may be encouraged. For example, unsupervised cluster analyses of severe asthmatics within the Severe Asthmatic Research Program (SARP) identified four distinct groups of patients with variable responses to treatment with corticosteroids[30]. Similarly, a topological data analysis of multi-omic data collected from severe asthmatics identified six distinct asthma endotypes[160]. The latter study highlights the benefits of performing integrated analyses using multiple datatypes in order to gain a greater understanding of disease mechanisms than what is possible with single data type (modality) analyses alone. This was further supported by a recent study which explored a potential methodology to integrate different data types to classify childhood asthmatics using supervised machine learning methods; the study demonstrated the predictive benefit offered by integrating multiple data modalities compared to using single modality data alone[161].

Although the majority of machine learning applications within the asthma field have focused on adult asthma, prediction models have also targeted paediatric populations. For example, Patel *et al.* used clinical and environmental data collected at emergency department triage from 29,392 patients to predict the need for hospitalisation required by children with asthma exacerbations[155]. Similarly, Goto *et al*. used clinical data from 52, 037 children admitted to an emergency department to predict two clinical outcomes at triage - the need for critical care or hospitalisation[162]. This study compared four machine learning algorithms against a conventional triage approach (reference model). The machine learning models demonstrated significantly superior ability to predict a child's need for hospitalisation compared to the reference model. They also offered better performance compared to the reference model to predict the need for critical care, but these improvements were not significantly different. Furthermore, using data from 112 patients from a hospital paediatric department in Greece, Chatzimichail *et al*. attempted to apply a number of feature selection methods and modelling algorithms to predict childhood asthma[163-166]. However, these studies remain as exploratory methodological studies due to the lack of independent replication.

1.3.8 The promise of explainable machine learning models

Despite the increasing adoption of complex machine learning methods to solve a variety of healthcare problems, their reputation as "black-box" algorithms have often posed a major hurdle for clinical application and utility. In an environment where decision making is critical and accountability is high, machine learning models assisting clinical decision-making need to be able to explain how predictions were made[167].

As explained in Chapter 1.3.1, complex machine learning models that may offer greater predictive performance are often less interpretable. Researchers are therefore faced with a trade-off between developing complex machine learning models that may offer superior performance and developing simpler models which offer poorer performance but are highly interpretable[167]. It is due to the importance of the latter that simpler regression and decision tree models have often shown to dominate the field of disease prediction.

Recently, novel approaches to increase the explainability of machine learning models have been proposed. For example, novel models, such as Generalized Additive Models plus Interactions

(GA2M), offering both high performance and interpretability, have been proposed[167]. Other approaches involve extracting explanations for “black-box” prediction models post-hoc using techniques, such as SHapley Additive exPlanations (SHAP)[168] and Locally Interpretable Model-agnostic Explanations (LIME)[169]. The benefit of these post-hoc techniques is that they can be used on any machine learning model[168], thus allowing researchers to address the issues of model interpretability and “user trust” whilst also taking advantage of the improved performance offered by complex models[167].

1.4 Aims

Current prediction models, across healthcare settings, are commonly based on traditional regression methods, due to their ease and high interpretability making them a popular choice among the clinical community. However, for the proportion of models which have undergone validation in independent populations, these models often demonstrate poor/modest generalisability. In part, this may be due to limitations inherent in the traditional statistical methods used. Machine learning approaches for prediction have been increasingly explored. Despite some conflicting studies assessing their benefit over traditional methods, numerous studies have shown machine learning methods to offer superior predictive capability compared to traditional regression methods. However, only a few studies have explored the application of machine learning for prediction in the context of childhood asthma.

Predicting a child’s risk of developing asthma is important to tailor asthma management in a bid to curb the progression and severity of the disease. Early prediction and effective intervention prior to disease development could also promote asthma prevention. However, the difficulty in predicting the development of childhood asthma stems from its complex pathophysiology and highly heterogeneous presentation, particularly in early childhood. Utilising information from a variety of available data modalities, such as questionnaires, clinical tests and -omic analyses, have already shown able to untangle clinical questions on the identification and categorisation of asthma phenotypes and endotypes in a hope to tailor future asthma management on an individual level[30,161]. Given the increasing availability of different data types emerging from both research and clinical settings, such integrative analyses may also promote more accurate predictions to be made in early life regarding the development of childhood asthma at school-age.

The Isle of Wight Birth Cohort (IOWBC) is the earliest European birth cohort established to study the natural history and development of asthma and other allergic diseases in early life[170]. With a rich collection of life-course data (detail in Chapter 2.1.1), the IOWBC offers the unique opportunity to develop prediction models for childhood asthma using a rich variety of clinical, environmental and genomic biomarker data collected in the early years of life.

Therefore, the aim of this thesis is to use methods of machine learning and data integration to develop two prediction models for childhood asthma development in the IOWBC. One model will aim to predict the risk of school-age asthma in early life, for the future purpose of asthma prevention, whilst the other will aim to predict the risk of school-age asthma by preschool-age, for the purpose of asthma management. To achieve this, the following objectives will be met:

1. Identification and critical evaluation of current risk models available for predicting childhood asthma;
2. Evaluation of the generalisability of existing childhood asthma prediction models within a single population;
3. Identification of readily available clinical features predictive of the development of school-age asthma;
4. Comparison of machine learning algorithms and independent selection of two optimised clinical prediction models for school-age asthma for wide-spread clinical use;
5. Generation and interpretation of personalised risk score probability estimates based on the chosen models;
6. Generation of polygenic and epigenetic risk scores for childhood asthma development;
7. Evaluation of any predictive improvement following the integration of the genomic risk scores with the clinical asthma prediction models.

Chapter 2 Methods

Detailed descriptions of the datasets and main methods used throughout this thesis to develop and independently validate the childhood asthma prediction models using machine learning and data integration processes are outlined in this chapter. Additional methods specifically used to address certain thesis aims are further described within the relevant chapters.

2.1 Datasets

2.1.1 Development dataset: Isle of Wight Birth Cohort (IOWBC)

The IOWBC, also known as the second generation (F_1) cohort, is based on the Isle of Wight (IoW), off the south coast of England, UK. It is a whole population, single-centre prospective cohort study established in 1989 to explore the natural history and development of asthma and other allergic diseases in early life[170]. Between 1989 and 1990, 1509 women gave birth to 1536 children on the Isle of Wight. From these births, 1456 children were recruited into the IOWBC study and followed up from birth and at 1, 2, 4, 10, 18 and 26 years, with high retention rates of 94.0%, 84.5%, 83.7%, 94.3%, 90.2% and 70.9%, respectively[170]. The IOWBC cohort comprises of individuals predominantly of Caucasian ethnicity (98%).

2.1.1.1 Clinical and environmental data

In the IOWBC, allergic disease and exposure-related data was collected through hospital records, physical examinations and study specific questionnaires. From the 10 year follow-up, questionnaires were standardised with the International Study of Asthma and Allergies in Childhood (ISAAC) questionnaire created in 1995[171].

At each follow-up, demographic and lifestyle information as well as pregnancy and birth characteristics, environmental exposures and indicators of asthma and allergy status were collected (total number of variables: at birth=70; 1-year=124; 2-year=110; 4-year=115; 10-year=306; 18-year=430; 26-year=460). Specifically, this included data on family history; gestational factors, breastfeeding and early life diet; household pets; exposure to tobacco smoking; housing characteristics; socioeconomic status as well as height, weight and BMI. Clinical symptoms for which data was collected included: wheeze, cough, nasal symptoms, nocturnal symptoms, chest infections, eczema, allergic rhinitis, food allergy and asthma. In addition, a skin prick test (SPT) was performed in infants with allergy-related symptoms at 1 and 2 years, and in all participants from the 4-year follow-up onwards. Sensitisation to the following 13 common inhaled and food allergens was assessed: house dust mite (*Dermatophagoides pteronyssinus*), grass pollen mix, tree pollen mix, cat and dog epithelia, *Alternaria alternata*, *Cladosporium herbarium*, milk, hen's egg, soya, cod, wheat, and peanut. A positive SPT was confirmed if the mean wheal diameter was at least 3mm greater than the negative control (physiologic saline). An individual was considered atopic following at least one positive SPT. Lung function tests, including spirometry and methacholine bronchial challenges, were performed from the 10-year follow-up. Bronchial challenge consisted of the serial administration of incremental doses of methacholine, from 0.0625mg/mL to 16mg/mL. In line with ATS guidelines, a child was deemed to have a positive test for bronchial hyper-responsiveness if <4.0mg/mL caused a 20% fall in FEV_1 (forced expiratory volume in 1 second) from the baseline FEV_1 value[14,172].

Candidate predictors considered for inclusion in the prediction models were shortlisted based on knowledge of risk factors associated with childhood asthma published in the literature (detailed in Chapter 1.1.6). The list of candidate predictors was further filtered to include only data which was available in the development and validation cohorts. Where applicable, information on these risk factors was extracted from the IOWBC across three distinct time-points: i) at birth, ii) in early life (combination of 1 and 2-year follow-up data) and iii) at preschool age (4-year follow-up data). In total, data for 54 candidate predictors were extracted from the IOWBC and considered during the development of the childhood asthma prediction models (Table A1).

For candidate predictors collected at the early-life time-point, categorical data was combined based on the highest level of exposure reported at either the 1-year or 2-year follow-up. Early life BMI was reported as BMI at the 1-year follow-up. Measures of child BMI (collected at 1-year and 4-years) were standardised against the British 1990 growth reference[173]. Furthermore, based on expert opinion and a significant chi-squared test of association (X^2= 37.55, p-value= 3.293e^{-10}, performed on individuals within the IoW 3rd Generation (F_2) cohort (n=181)), data on “frequent wheeze” was used as a surrogate variable to account for the commonly evaluated predictor, “wheeze apart from cold”. Based on a cluster analysis using information on i) family income at 10 years of age; ii) the number of children in the index child’s bedroom; and iii) the British socioeconomic classes, socioeconomic status in the IOWBC was categorised into five distinct clusters[174].

2.1.1.2 Genotype data

Blood samples were collected from the child at multiple time-points - from heel pricks at 7 days of age (collected on Guthrie cards) and in later childhood at 10, 18 and 26 years. DNA from the peripheral blood samples of 1067 individuals in the IOWBC were isolated and underwent genome-wide genotyping using the Illumina InfiniumOmni2.5-8v1.3 microarray. Blood samples collected at the 18 year time-point were used for genotyping; where unavailable, blood samples collected at age 7 days, 10 years or 26 years were used. Standard quality control for genome-wide association studies (GWAS) had been performed to exclude samples with low call rate (<97%) and SNPs with call rate <95%, minor allele frequency (MAF) <0.005 and significant deviation from Hardy-Weinberg equilibrium (p-value <1x10^{-8}) prior to imputation. Alleles had also been updated to match the direction (forward) and coordinates of the reference dataset, GRCh37[175]. Data were pre-phased (EAGLE2)[176] and imputed (PBWT)[177] using the Sanger Imputation Services (Oxford, UK).

Further quality control was performed to prepare the imputed genotype data for the analyses conducted in this thesis. This included the retention of data with an imputation quality >80%. SNPs were again filtered to exclude those with call rate <95%, MAF<0.01 and significant deviation from Hardy-Weinberg equilibrium (p-value <1x10^{-6}). Samples were further filtered to remove those with call-rate <97%, extreme heterozygosity (±3SD of the mean F-coefficient) and gender mismatch. One individual of each related pair (3rd degree relations or closer, pi-hat ≥0.125) was also excluded. Population structure was assessed by principal component analysis (PCA), comparing the IOWBC with the European descent (CEU), Yoruba (YRI), Hans Chinese (CHB) and Japanese (JPT) HapMap3 reference populations[178]. Non-European individuals were excluded based on a visual inspection of the PCA plot (Figure A1). A final dataset of 977 individuals with genotype data for 7,236,427 SNPs was retained for downstream analyses.

In addition, five candidate variants of the filaggrin gene (R501X, 2282del, S3247X, 3702delG and R2447X) had previously undergone genotyping in the IOWBC using GoldenGate Genotyping Assays (Illumina, Inc, SanDiegom CA) on the BeadXpress Veracode platform per Illumina's protocol[179]. In brief, 1,248 blood samples (from 1,211 individuals and 37 replicates) were fragmented, hybridised to allele-specific primer sets and subject to extension/ligation reactions. Samples were sourced from a similar mixture of time points as described for the genome-wide genotype data. Samples were then hybridised to the Veracode bead pool for processing by the

BeadXpress reader. Allele determination was based on a GenCall score >0.25. Scores below this quality threshold were deemed “no calls”.

For downstream analyses conducted in this thesis, quality-controlled genotype data for the R501X variant of the filaggrin gene (n=924) was added to the quality controlled imputed genome-wide genotype profiles detailed above, resulting in a final dataset of 924 genotype profiles consisting of 7,236,428 SNPs.

2.1.1.3 Methylation data

Genome-wide DNA methylation data was measured at birth from blood samples collected on Guthrie cards (n=885). DNA from Guthrie cards was extracted using the Gensolve kit, following the procedure described by Beyan *et al.*[180]. 500ng of isolated genomic DNA from each sample was then bisulphite-treated using the EZ 96-DNA methylation kit (Zymo Research, Irvine, CA, USA). This process facilitates the deamination of unmethylated cytosines (converting them to thymine) at CpG islands whilst leaving methylated cytosines unchanged. DNA methylation profiling was then performed using the Illumina Infinium MethylationEPIC BeadChips following the manufacturer’s standard protocol. In this process, the DNA methylation levels at 863,904 CpG sites were estimated as beta values - the ratio of the methylated probe intensity and the overall intensity (Equation 2.1). DNA methylation beta values range from 0 (completely unmethylated CpG) to 1 (completely methylated CpG). Due to funding limitations, Guthrie DNA methylation data was collected and profiled (sent to the same service provider) in seven batches.

$$\beta = \frac{Methylated\ signal\ intensity}{Unmethylated\ signal\ intesity + methylated\ signal\ intensity + 100}$$

Equation 2.1 DNA methylation beta value estimation

Next, DNA methylation data underwent a number of pre-processing steps. Beta values were normalised using the CPACOR method[181]. Illumina Background Correction was applied to intensity values prior to the exclusion of CpGs with intensity values with detection p-values ≥10^{-16} and samples with call-rate <95%. Using the minfi package[182], gender was inferred based on the difference in median total intensity of CpGs on the X and Y chromosomes. Individuals whose predicted gender directly contradicted their reported gender, as well as those who deviated ±4SD from the main gender clusters, were excluded. One individual of each set of repeat samples and related pair of individuals (3rd degree relations or closer, pi-hat ≥0.125) were retained. Quantile normalisation was applied to intensity values using the DASEN method[183], incorporating control probe adjustment and global correction reduction. As DNA methylation was measured in seven batches, batch effects were removed using ComBat (sva package)[184]. SNP-associated and cross-hybridised probes were removed[185]. A total of 765 individuals with DNA methylation profiles consisting of 694,571 CpGs were retained for further analyses.

2.1.2 Replication dataset: Manchester Asthma and Allergy Study (MAAS)

The Manchester Asthma and Allergy Study (MAAS) is an unselected birth cohort which was established to study the development of asthma and other atopic disorders in childhood[186]. Participants were recruited into the study from 50 square miles of South Manchester and Cheshire (within the maternity catchment area of the Wythenshawe and Stepping Hill Hospitals). Between 1995 and 1997, 1211 women (≤10 weeks pregnant) were recruited into the study and 1184 children were subsequently followed up at 1, 3, 5, 8, 11, 13-16 and 18 years. The MAAS cohort consists of a stable mixed urban-rural population (~89% Caucasian). The 13-16 year follow-up questionnaires were harmonised with the Isle of Wight cohort as part of the STELAR (Study Team for Early Life Asthma Research) Consortium[187].

2.1.2.1 Clinical and environmental data

Medical records and validated questionnaires were used to collect data on family history, clinical symptoms of asthma and allergy as well as environmental exposures[186,188]. SPT for house dust mite (*Dermatophagoides pteronyssinus*), cat, dog, grass pollen, moulds, milk, and egg were performed from the 3-year follow-up onwards; tree pollen and peanut allergens were also tested from the 8 year follow-up onwards. A mean wheal diameter at least 3mm greater than the negative control (physiologic saline) was used to confirm a positive SPT. An individual was considered atopic following at least one positive SPT. Lung function tests were also performed from the three-year follow-up. From the 8-year follow-up, methacholine bronchial challenges (0.0625-16.0 mg/mL) were performed using a 5-step protocol in line with ATS guidelines. Bronchial hyper-responsiveness was confirmed by a 20% fall in FEV_1 from the baseline measurement.

In accordance with the candidate predictors considered in the IOWBC, information in MAAS was assessed at birth (recruitment), in early life (combination of 1 and 3 year follow-ups) and at preschool age (5 year follow-up).

2.1.2.2 Genotype data

DNA samples from 919 individuals underwent genome-wide genotyping using the Illumina 610 quad chip, with genotypes called using the Illumina GenCall application following the manufacturer's instructions. Prior to imputation, data was quality controlled to exclude SNPs with call rate <95%, MAF<0.005 and significant deviation from Hardy-Weinberg equilibrium (p-value $<3x10^{-8}$). Alleles had also been updated to match the direction (forward) and coordinates of reference dataset, GRCh37[175]. Samples with low call-rate (<97%), outlier autosomal heterozygosity, gender mismatch and non-European ethnicity were excluded. One individual from each pair of siblings or cryptic relations was also excluded. Genotypes were then imputed using IMPUTE version 2.1.2, with the 1000 Genomes and HapMap Phase 3 reference genotypes[178].

For analyses performed as part of this thesis, imputed genotype data was further quality controlled to exclude SNPs with low imputation quality (INFO<0.80), MAF<0.01 and significant deviations from Hardy–Weinberg equilibrium (p-value $\leq 1 \times 10^{-8}$). A final dataset of 852 individuals with genotype profiles of 7,353,200 SNPs were retained for downstream analysis.

2.2 Prediction outcome: School-age asthma

The prediction outcome of interest in this thesis is the development of school-age asthma. In the developmental dataset (IOWBC), school-age asthma was defined by a combination of a doctor diagnosis of asthma ever and at least one episode of wheezing or use of asthma medication in the last 12 months[170]. For each analysis, all participants with a reported asthma outcome at age 10 were considered (n=1368, asthma prevalence=14.6%).

To validate the prediction models in MAAS, the outcome of school-age asthma was constructed to fully correspond with the definition used in the developmental dataset. Data on school-age asthma was assessed at both 8 (n=1018, asthma prevalence=14.1%) and 11 years (n=898, asthma prevalence=12.9%) in MAAS.

2.3 Machine learning approaches

2.3.1 Feature selection

As previously discussed, feature selection methods aim to identify a subset of the most predictive features, reducing model dimensionality and noise, in order to improve computational demand and potentially improve prediction accuracy. Two feature selection methods, Recursive Feature Elimination (RFE) and Boruta, were compared to identify a subset of predictors with high classification accuracy from the list of 54 candidate predictors. Both are wrapper methods which utilise the random forest algorithm. A variation of the random forest algorithm (balanced random forest)[189], which randomly under-samples the majority class in each bootstrap, was used to account for the class imbalance present in the dataset.

2.3.1.1 Recursive feature elimination

Recursive Feature Elimination (RFE) is an example of a backward elimination process. RFE was initially developed as a wrapper method utilising the SVM algorithm[190]. Given its in-built estimation of feature importance and easy application without the need for hyperparameter tuning, the random forest algorithm has also been used for RFE and has shown competitive performance[191].

RFE aims to identify a subset of important features through an iterative process of evaluating the relative importance of features and removing those deemed least important. Initially, all candidate features are used to train the wrapper algorithm of choice (e.g. random forest). Features are each assigned a weighting and ranked based on a specified importance criterion. The lowest ranking feature is removed from the pool of candidate features. This process is repeated, whereby the remaining features are used to retrain the random forest algorithm and update the feature importance ranking. In this manner, the lowest ranking features are recursively eliminated until only a single feature remains. At each elimination step, one or multiple features can be removed at a time[190,191].

RFE using the random forest algorithm can be applied within a *k*-fold cross validation framework (RFECV), whereby, for each split, RFE is performed on (*k-1*) partitions and the model's performance is tested on the k^{th} partition. The optimal subset of features can then be identified by the subset demonstrating the best cross-validation performance (Pseudocode 1).

A key aspect of RFE is that the subset of selected features does not necessarily consist of the features with the highest individual importance rankings; instead the selected features have the top ranking when considered as part of the feature subset[190]. Hence, this method aims to overcome the limitations of univariate and filter methods that consider features independently, ignoring their potential inter-relationships. In addition, due to re-evaluating feature importance after each elimination step, RFE has been suggested to be a beneficial feature selection method in the presence of highly correlated features[191]. However, this has not shown to extend well in highly dimensional datasets with a large number of correlated features[191,192].

2.3.1.2 Boruta

The Boruta algorithm aims to identify a subset of relevant features among the original candidate feature list by evaluating the importance of each feature compared to random variables. By comparing original candidate features against random variables, the idea is to account for any correlations between the trees in a random forest that may artificially increase the true importance of features. Truly relevant features are expected to have higher importance than any randomly generated variable.

To implement Boruta, the feature space is extended to contain both the original features and a set of shadow features (shuffled replicates of the original features). The importance of each feature is evaluated as a Z-score, computed as the average loss of accuracy divided by the standard deviation across the trees in a random forest. A hit is assigned to each original feature that has a higher Z-score than the maximal Z-score reported amongst all of the shadow features. This process is repeated for a number of iterations. A statistical two-sided test of equality is then performed to evaluate whether the observed number of hits for a feature was higher than expected over the iterations. Features that demonstrate high feature importance (Z-score) more times than expected are considered important and included in the feature selection subset (Pseudocode 2)[193,194].

2.3.2 Model Development

To identify the best algorithm for this classification problem, the predictive performance of eight supervised machine learning classifiers were compared: support vector machines (SVM) with three different kernel functions (linear, radial basis function and polynomial), decision tree, random forest, naive Bayes classifier, multilayer perceptron (MLP), and K-Nearest Neighbours (KNN).

2.3.2.1 Support vector machine

SVM is an example of a highly flexible classification algorithm. SVMs aim to construct a separating hyperplane between outcome classes (Figure 2.1)[118]. The hyperplane is constructed with the aim of maximizing the separation between the data points closest to the decision boundary (support vectors). In theory, the larger the margin between the decision boundary and the support vectors, the better the classifier. However, data is often not linearly separable. In such cases, a soft margin, which allows for a degree of misclassification of the support vectors, can be applied to obtain the best classification[195,196]. The addition of a soft-margin has also shown to be beneficial even when data is linearly separable. In the construction of the soft margin, slack variables (ξ) for each example, are used to assign a degree of error to the classifications, allowing examples to fall within the margin (ξ between 0-1) or be misclassified (ξ>1). However, to control the use of slack variables, a cost parameter is introduced with the aim of maximising the margin whilst minimising the amount of slack[196]. The cost parameter (*C*) is a regularization term that assigns a penalty to misclassifications. This regularisation term is a hyperparameter of the SVM, which can be tuned in order to provide the optimal classification based on a given dataset. Classifiers with larger values of *C* incur higher penalties for misclassifications and therefore have smaller margins (often resulting in the risk of overfitting)[119].

Particularly for non-linearly separable data, the accuracy of a classifier can be improved with a non-linear decision boundary. This flexibility of the decision boundary can be achieved using the kernel trick, whereby the data in the original feature space is mapped onto a higher dimensional space, which in turn, enables a linear hyperplane to be constructed. When projected back onto the original feature space, the decision boundary appears non-linear[119,196]. Different kernels such as the radial basis function (RBF) or polynomial function can be used for this purpose. The RBF and polynomial kernels specify a gamma (γ) hyperparameter, which is a scaler property to determine the influence of each data point on the classification. Larger values for γ encourage the classifier to overfit onto the training data. Whilst the RBF kernel projects data onto infinite dimensions, the polynomial kernel determines dimensionality by a degrees (*d*) hyperparameter. Large values of *d* increase the flexibility of the classifier but can lead to overfitting[119,196].

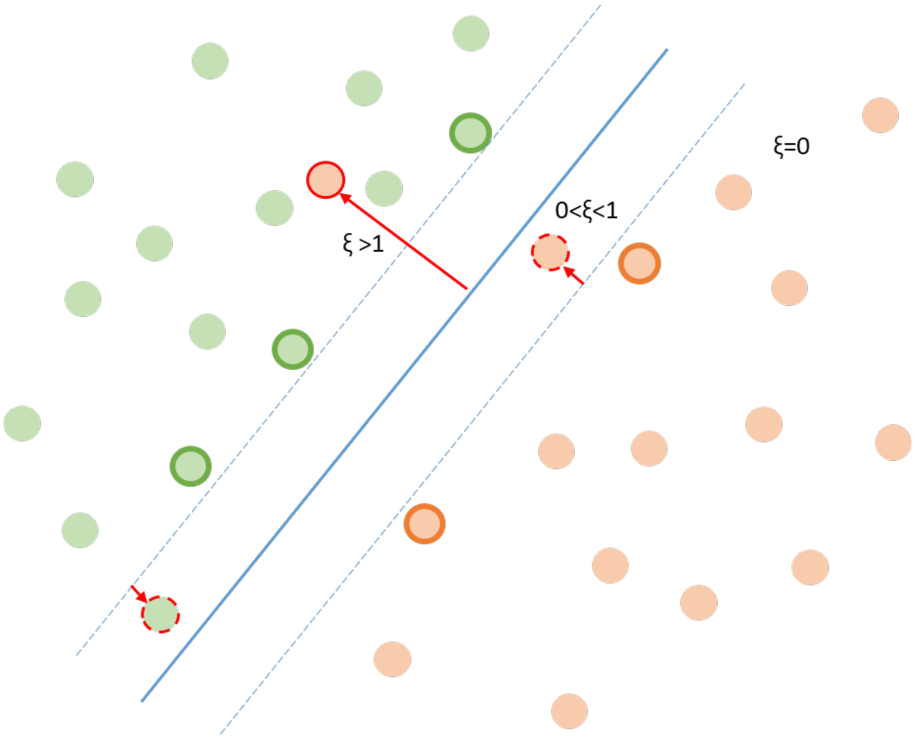


Figure 2.1 Schematic of the support vector machine classifier

The thick blue line indicates the decision boundary which has been constructed to separate examples which belong to two different classes (green and orange circles). The dotted blue lines depict the margin which has been constructed to maximise the separation between the decision boundary and the support vectors (green and orange circles with bolded outlines). With a soft margin, the size of the margin is increased by allowing some examples to be within the margin (circles outlined with red dashes) or be misclassified (circles with a solid red outline). Slack variables are assigned to examples found within the margin or those which are misclassified (*ξ*). A cost function (*C*) is used to penalise misclassifications, and is used as a regularisation term to maximise the size of the margin whilst minimising the amount of slack. Figure reproduced based on Ben-Hur *et al.*[196] (copyright license: CC-BY).

In this comparative analysis of machine learning models, the linear, radial basis function (RBF), and polynomial kernels were each used to construct three different SVM classifiers.

2.3.2.2 Decision tree

The decision tree is a highly interpretable machine learning algorithm which aims to stratify the predictor space using a number of splitting rules. Starting at the top (root node) of the tree, each point at which the predictor space is stratified by a variable is referred to as an internal node. The final nodes at the bottom of the tree, at which no further separations are made, are referred to as leaves or terminal nodes and provide the final classification (Figure 2.2)[118].

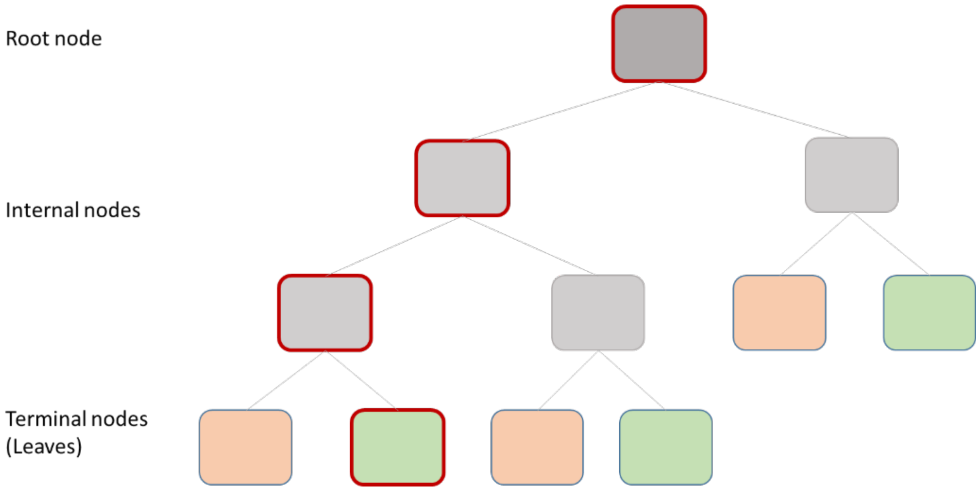


Figure 2.2 Schematic of the decision tree algorithm

Starting at the top (root node), a decision tree is split into branches (grey lines) based on the feature that generates the best class separation. At the end of each branch (internal nodes), another split occurs based on the feature offering the best separation out of the remaining features. This process is repeated until no further splitting can occur or the node is pure (leaf/ terminal nodes). The nodes outlined in red provide an example for how an example may be classified into its class (orange or green). Figure produced based on James *et al.*[118].

The splitting at each internal node is determined using a top-down greedy search strategy, whereby the features that generate the best separation of the outcome classes are chosen at each node. A number of criteria can be used to define the separation at each node, with the Gini index or entropy being the most commonly used metrics. The degree to which the outcome classes are separated at a node is referred to as the node purity; a node with perfect separation of the outcome classes is considered a pure node. The node purity is tested for all remaining features at each subsequent node along the branch. In this way, the features demonstrating the best separation are found higher up in the tree structure. For continuous features, the splitting threshold is determined using recursive binary splitting. This process assesses all possible thresholds to identify the cut-off value offering the purest separation. Categorical features are split by their defined categories or, if ordinal, can be stratified into two distinct categories similar to continuous features. Tree-based algorithms work well with categorical features but tend to favour features with multiple levels, due to their increased chance of identifying a good separation of the data[119,197]. When the separation of the previous node cannot be improved upon further, the stratification process stops, and the node becomes a leaf node.

The splitting criterion determining the construction of the decision tree algorithm is highly dependent on the training data. As a result, these models are considered highly unstable and often generalise poorly. The stratification stopping criterion used in the construction of the decision tree aims to reduce this instability. A number of stopping criteria, such as attaining node purity, reaching a maximum tree depth or having a minimum number of cases in the node can be specified. Additionally, whilst the growth of the tree will automatically stop when no further improvement in the splitting criteria can be obtained, it is possible to specify a certain criteria upon which splitting is allowed (e.g. a minimum of *n* samples are required at the node for a further split to occur). These, alongside other pruning methods, aim to reduce the complexity of the model in order to prevent overfitting and reduce the variance of the model[118,197].

2.3.2.3 Random forest

The random forest algorithm is an example of an ensemble classifier that aggregates the decisions of multiple decision trees in an aim to reduce the variance observed by individual decision trees[198] (Figure 2.3).

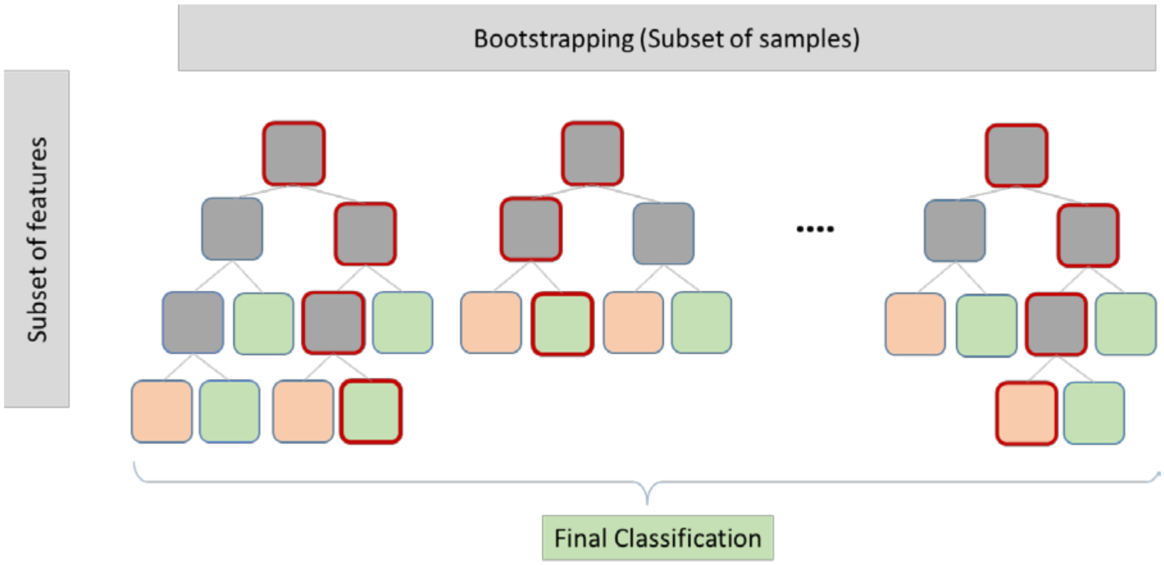


Figure 2.3 Schematic of the random forest algorithm

The random forest model aggregates the decisions of multiple decision trees to make a final classification (bagging). To construct each decision tree, only a subset of features and a subset of samples (bootstrapping with replacement) are used. Each decision tree will have low bias and high variance but the overall variance of the ensemble model will be low. Figure produced based on James *et al.*[118].

For the development of each decision tree, a bootstrapped dataset the same size as the original dataset is created using a resampling process whereby samples are randomly selected with replacement from the original dataset[118]. Unlike the decision tree algorithm, for each tree in the random forest, only a random subset of variables is considered for stratifying the predictor space at each internal node in order to reduce the correlation between the trees. Although each decision tree will still have low bias and high variance, the process of bootstrapping and aggregating decisions across multiple trees to make a final classification (known as bagging) aims to offer predictions with low variance and high accuracy[118,198]. Hyperparameters, including those of a decision tree algorithm (detailed in Chapter 2.3.2.2), can be tuned in the construction of a random forest model to maximise performance.

2.3.2.4 Naïve Bayes classifier

Based on conditional probability (Equation 2.2), the naïve Bayes algorithm is one of the simplest supervised machine learning algorithms. It is underpinned by the assumption that each feature is independent of the others in determining the outcome class. The implementation of the naïve

Bayes algorithm requires assumptions to be made on the prior probability distribution of each class. Commonly, a Gaussian and multinomial/Bernoulli distribution is assumed for continuous and categorical features, respectively[118].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Equation 2.2 Conditional probability estimation underpinning the Naive Bayes algorithm

Bayes theorem calculates the conditional class probability given a predictor, P(A|B).
P(B|A) is the probability of the predictor given the class (likelihood).
P(A) is the prior probability of the class.
P(B) is the prior probability of the predictor.

2.3.2.5 Multilayer perceptron

The multilayer perception (MLP) is categorised as a simple feed-forward artificial neural network. Neural networks are particularly well-suited to distinguish non-linearly separable data through a network of interconnected nodes[119,199]. The architecture of the MLP consists of a minimum of three layers; an input layer, at least one hidden layer and an output layer (Figure 2.4). Within the input layer, the number of neurons is equivalent to the number of features. The input neurons serve only to pass the input vector to the nodes in the next layer. Each neuron is fully connected, with an assigned weight, to each neuron in the next layer, for all hidden layers. The final classification of the outcome is determined based on the output node with the greatest signal.

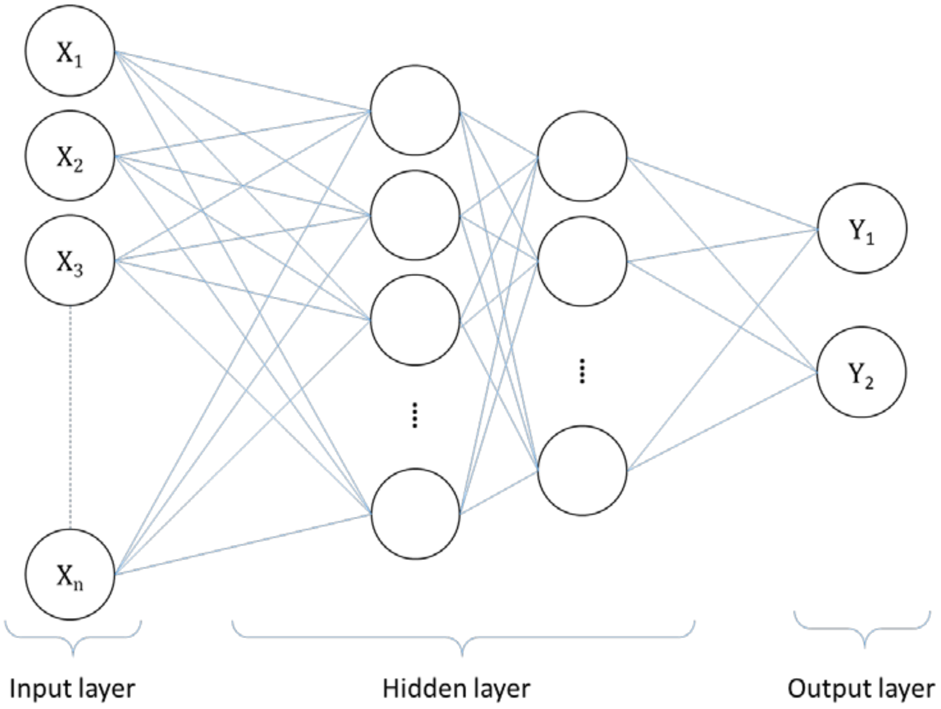


Figure 2.4 Schematic of the multilayer perceptron algorithm

A multilayer perceptron consists of an input layer, with a neuron for each input feature (*X_1* to *X_n*); an output layer with a neuron for each outcome class (e.g. *Y_1* and *Y_2* for a two-class outcome); and a hidden layer. The hidden layer can comprise of one or more layers of neurons. Each neuron is fully connected to the neurons in the next layer. Figure reproduced based on Gardner *et al.*[199], with permission from Elsevier.

The strength of the signal at each node is determined by the sum of the weights inputting the output node, modified by a non-linear activation function[118]. During the training process, the connection weights between neurons are optimised to minimize the error of the output layer through a process of backpropagation. The error of the output layer for each combination of weights can be considered as an error surface. The backpropagation algorithm aims to locate the global minimum of the error surface through gradient descent. In this process, small weights are randomly assigned to the neurons and the local gradient of the error surface is calculated. The weights are adjusted in the direction of the steepest local gradient and the local gradient is recalculated. This is repeated until the adjusted weights converge to the global minimum[199].

However, as well as the global minimum, the error surface can have multiple local minima. Depending on the random starting point of the back-propagation algorithm, there is potential for the algorithm to get trapped at a local minimum, unable to find the global minimum. To address this, there are two main tuning hyperparameters of the MLP algorithm, the learning rate and momentum. The learning rate addresses the step size taken during gradient descent. Step sizes which are too large can result in erratic changes in the weights due to large variation in the local gradients calculated. In contrast, finding the global minimum using small step sizes can be time-consuming. The momentum parameter provides assistance in the gradient descent if the algorithm gets stuck at a local minimum by adding a proportion of the previous weight-change to the change in the current weight[199].

2.3.2.6 K-nearest neighbours

K-nearest neighbours (KNN) is an instance-based learning algorithm. It aims to classify an unknown data point based on the *k*-data points in closest proximity (nearest neighbours) for which the class labels are known[118] (Figure 2.5). First, the conditional probability of the unknown data point belonging to each class is calculated. The calculated probability is then used to assign the classification to the class with the largest probability (i.e. the modal class of the *k*-nearest neighbours). As a result, the classifications made by this algorithm are highly dependent on the number of neighbours considered. Whilst very small values of *k* can be subject to noise and affected by outliers, larger values of *k* can bias classifications against the minority class[118].

The nearest neighbours used to classify a given data point are determined using a distance measure, commonly calculated by either Euclidean or Manhattan distance. The influence of each neighbour on the classification can be uniform for all neighbours or weighted by the inverse of their distance[118].

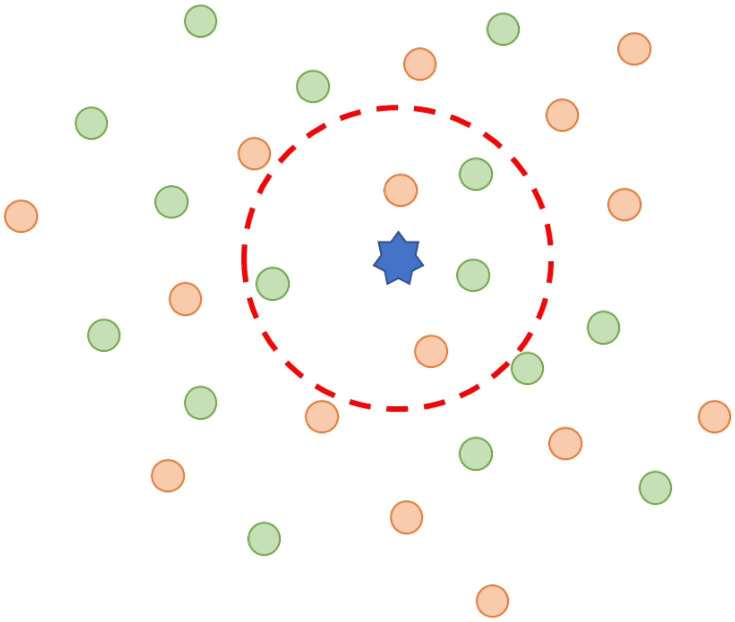


Figure 2.5 Schematic of the K-Nearest Neighbours algorithm

An unknown example (blue star) is classified based on the *k*-nearest examples. The dashed red circle identified the nearest neighbours (*k*=5) used to classify the unknown example. In this example, when *k*=5, the example would be classified as belonging to the green class. Figure reproduced based on James *et al.*[118], with permission from Springer Nature.

2.3.3 Hyperparameter tuning

Each of the machine learning algorithms described have a set of hyperparameters which can be tuned to promote optimal classifications based on a given training dataset. Different algorithms have different tuning parameters, but not all the parameters are important for tuning purposes[200].

Grid search is a hyperparameter search strategy that conducts an exhaustive search of each combination of hyperparameters to identify the optimal set of hyperparameters based on a specified performance measure within a cross-validation framework. Grid search is a popular search strategy but can be computationally expensive depending on the algorithm and the size of the hyperparameter space explored. In addition, it is possible for the performance of the grid search to be compromised by the consideration of a large number of hyperparameters[200].

In contrast, a random search strategy involves the random selection of a specified number of hyperparameter combinations to be evaluated. As this method is not an exhaustive search across all possible hyperparameter combinations, it is substantially faster and less computationally demanding compared to grid search. However, there is potential for the random search to miss the optimal hyperparameter set. Yet, studies indicate that random search is beneficial for the quick evaluation of a large hyperparameter search space with a focus on important tuning parameters, and is often able to identify the optimal hyperparameter set[200].

When large parameter spaces are being evaluated, it is suggested that a dual search strategy may be used to combine the advantages of both the random and grid search approaches. In this dual search strategy, a random search can first be conducted to quickly evaluate and narrow down a large hyperparameter search space. An exhaustive grid search can then be employed across the condensed search space to definitively identify the best hyperparameter set[200].

To tune the hyperparameters of the eight machine learning algorithms compared in this thesis, a grid search was used (Table A2). However, due to the large hyperparameter space considered for the models developed using the SVM algorithm, the dual search strategy was used to reduce computation time. Other than distinguishing between the continuous and categorical variables, the naïve Bayes model did not require any hyperparameters to be tuned. Hyperparameter tuning was only performed during model training; it was not performed on the random forest algorithm during feature selection as the default parameters are claimed to offer good predictive accuracy when applied across many problem settings.

2.3.4 Imputation

There are a number of methods available for imputing missing data[143]. A comparison of imputation methods addressing the problem of missing medical data identified missForest and MICE as the overall two best imputation methods; these methods demonstrated the lowest imputation error in two large datasets containing both continuous and categorical data, across different simulated proportions of missingness[143]. Hence, these two methods were compared in this thesis to impute missing data among the predictors identified from the feature selection.

2.3.4.1 MissForest

MissForest imputation is underpinned by the random forest algorithm and is a type of single imputation[148]. Initially, all missing values are assigned a placeholder value based on an imputation method such as mean imputation. The variables in the dataset are then ordered in ascending order of the proportion of missing data. For each variable with missing data, x_i, the dataset can be separated into: y_{obs} = examples with observed values for x_i; y_{mis} = examples with missing values for x_i; x_{obs} = the remaining variables for examples with observed values for x_i; and x_{mis} = the remaining variables for examples with missing values for x_i.

Starting with the variable with the lowest proportion of missing data, a random forest algorithm is trained using the features, X_{obs}, and outcome, y_{obs}. The trained model is then applied to the features, x_{mis}, to predict values for y_{mis}. The training and prediction steps are conducted in a cyclical manner for each variable with missing values. Once all missing values have been imputed, this whole process is iterated until a specified stopping criterion has been met. The stopping criterion is defined as the first point at which the difference between the newly imputed data and the imputed dataset generated from the previous iteration increases for both continuous and categorical variable types. The aim of the stopping criterion is to ensure that the properties of the new imputed dataset does not vary greatly from the non-imputed dataset. Once the stopping criterion has been met, a single imputed data matrix is returned for subsequent analyses[148].

2.3.4.2 Multivariate Imputation by Chain Equations

Multivariate Imputation by Chain Equation (MICE) is a type of multiple imputation used under the assumption that data is MAR. Unlike single imputation methods, which generate a single estimation for each missing value, multiple imputation utilises the distribution of the observed data in order to suggest multiple estimates for each missing value. By generating a set of plausible estimates, multiple imputation aims to account for the statistical uncertainty associated with the imputation[146].

MICE can perform imputation on datasets containing variables of mixed datatypes. For each variable, a different imputation model can be used depending on its datatype. In this thesis, the recommended imputation models were used – 'norm', a Bayesian linear regression model for numerical data; 'logreg', a logistic regression for binary data; and 'polyr', a proportional odds model for ordinal data[201]. In the implementation of MICE, all missing values are initially assigned a placeholder value based on mean imputation or random sampling (with replacement) of the observed data for each variable. For the first variable with missing data, x_1, the placeholder values are removed and x_1 is regressed on the remaining variables [x_2, x_3, …, x_i]. The regression is limited to only those examples for which x_1 was observed. The missing values for x_1 are then predicted from the posterior predictive distribution generated by the imputation model. This process is repeated for the remaining missing variables, where for example, x_2 is regressed on the remaining variables ([x_3, x_4, …, x_i] and the newly imputed variable (X_1)), again, limited to examples with observed data for x_2. Once all variables with missing data have been imputed, one cycle is complete. Numerous cycles are performed in order to converge the distribution parameters of each variable and create a single dataset of stable imputation estimates. To generate multiple (*m*) imputed datasets, this entire process is repeated *m* times[145,146].

Following the imputation stage and the generation of multiple imputed datasets, subsequent analyses should be conducted on each of the *m*-imputed datasets and the results should be pooled. The pooled results provide estimates with confidence intervals, addressing the statistical uncertainty of the imputation[145,146]. However, due to the need to tune each of the machine learning algorithm to establish a single model with a single set of tuned parameters, a single imputed dataset was required for model development. To form a single imputed dataset, the imputed values generated across the *m*-imputed datasets (*m*=5) were averaged, with the mean and modal imputed values taken for the continuous and categorical variables, respectively.

2.3.5 Resampling

There are numerous methods for handling class imbalances within a dataset[134]. Synthetic data generation methods such as SMOTE have demonstrated superior performance over random oversampling sampling and complete data analyses[136]. Numerous improvements of SMOTE, such as ADASYN, have been suggested in an attempt to tailor the oversampling procedure towards generating examples of the minority class which may promote a more accurate definition of the classification decision boundary[134,135,138]. Therefore, ADASYN was used in this study to help improve the class imbalance of the data.

2.3.5.1 Oversampling: Adaptive synthetic sampling

ADAptive SYNthetic (ADASYN) sampling is an example of a synthetic data generation approach based on the KNN algorithm. However, rather than randomly oversampling examples of the minority class, ADASYN prioritises the generation of difficult to classify examples of the minority class. To facilitate this, the synthetic examples generated through ADASYN are informed by a density distribution of weights for examples belonging to the minority class. The weight assigned to each example is determined by the ratio of examples belonging to the minority class in its *k*-nearest neighbours. These weights correspond to the learning difficulty of each example and subsequently determines the number of synthetic examples of the minority class that needs to be generated. For example, a difficult to classify example of the minority class (i.e. one that is similar to examples of the majority class) will have a small ratio of minority examples within its *k*-nearest neighbours, therefore will have a large weight. A greater number of synthetic examples will be generated based on this minority example. As a result, the learning model will have a greater opportunity to learn from difficult to classify examples of the minority class in addition to reducing the bias of the model by correcting for the class imbalance[138].

ADASYN can specify the construction of datasets with varying degrees of class balance. In datasets with a large class imbalance, oversampling the minority class can result in the construction of a dataset with a large proportion of synthetic data. To address this issue, different degrees of oversampling was performed in this thesis – the minority class was oversampled by 25%, 50%, 100%, 150%, 200%, 250% and 300% (resulting in up to 75% of the minority class consisting of synthetic data).

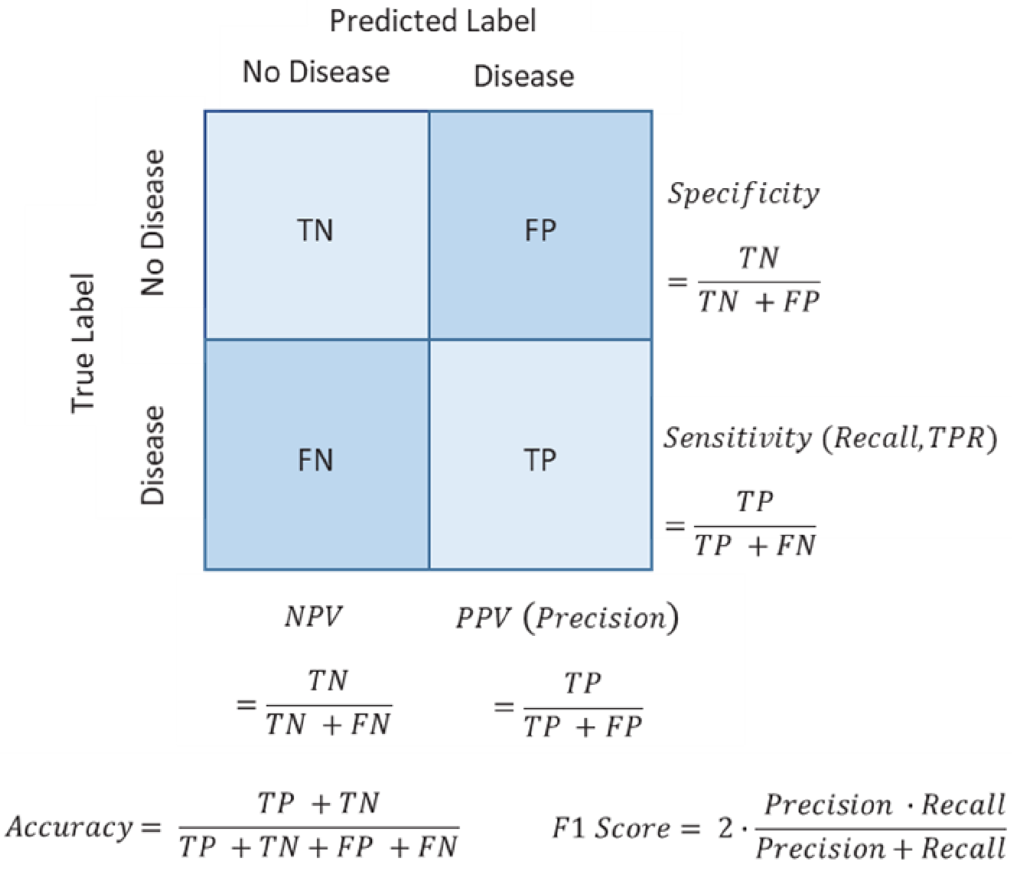
2.3.5.2 Random undersampling

Alongside oversampling the minority class, undersampling is another approach to balance the class proportions of a dataset. Undersampling involves the exclusion of examples of the majority class. In this thesis, examples of the majority class (non-asthmatic individuals) were randomly excluded to achieve a training dataset with 1:1 class ratio.

2.3.6 Predictive performance measures

Predictions made by a prediction model can be categorised into four main groups - true positive (TP) – correctly predicting those with the disease as having the disease; false negative (FN) – incorrectly predicting those with the disease as being disease-free; true negative (TN) – correctly predicting those without the disease as being disease-free; and false positive (FP) – incorrectly predicting those without the disease as having the disease. These results from a prediction model can be summarised in a confusion matrix and can be used to calculate numerous metrics to evaluate the predictive performance of the model (Figure 2.6)[202,203].

The accuracy of a model is defined by the proportion of correctly made predictions. Sensitivity, also termed recall or the true positive rate (TPR), is the proportion of individuals with the disease who are correctly predicted to have the disease. The positive predictive value (PPV), also termed precision, refers to the proportion of individuals correctly identified with the disease out of the total number with a positive prediction. Specificity refers to the proportion of individuals without the disease who are correctly predicted as being disease-free. The false positive rate (FPR) is the proportion of individuals incorrectly predicted to have the disease, calculated as (1-specificity). The negative predictive value (NPV) measures the proportion of individuals correctly predicted as being disease-free out of the total number with a negative prediction. A good predictive model should ideally have high sensitivity and high specificity. However, with the optimisation of either parameter promoting misclassification, there is an unavoidable trade-off between sensitivity and specificity.



		Predicted Label	
		No Disease	Disease
True Label	No Disease	TN	FP
	Disease	FN	TP

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity\ (Recall, TPR) = \frac{TP}{TP + FN}$$

$$NPV = \frac{TN}{TN + FN} \qquad PPV\ (Precision) = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Figure 2.6 Schematic of a confusion matrix and formula for the main metrics used to evaluate model performance

Figure produced based on James *et al.*[118].

A Receiver Operating Characteristic (ROC) curve, a plot of the TPR against FPR across all outcome probability classification threshold cut-offs, graphically demonstrates the discriminative performance of a model (Figure 2.7). The area under the ROC curve (AUC) is one of the most widely reported performance metric used to evaluate and compare prediction models[204]. The AUC ranges between 0 and 1; models with an AUC of 0.5 are considered to be no better than a random guess. The higher the AUC, the better the predictive performance of a model, with a model with an AUC of 1 considered to have perfect predictive performance. Models with a lower AUC (less than 0.5 and nearing zero) are considered to have poor predictive performance[117]. Whilst confusion matrices to assess model performance can be evaluated at any threshold cut-offs, it is common for performance to be reported at a classification threshold of 0.5 or at the threshold which maximises model informedness, the Youden's index (Figure 2.7)[204,205]. The Brier score, which measures the mean squared difference between the predicted probability of the outcome and the observed outcome, also evaluates overall model performance by considering both model discrimination and calibration characteristics; the score ranges from 0 to 1, with a lower Brier score indicative of a better performing model[205].

In some instances, the misclassification of diseased individuals as disease-free is preferred over unnecessarily exposing healthy individuals to treatments with potentially severe adverse effects. In such situations, models should have a low negative likelihood ratio (LR-), the probability of a false negative prediction against a true negative prediction. Conversely, when the benefits of treatment outweigh their potential risks, it may be preferred that a predictive model favours to rule in the disease. A high positive likelihood ratio (LR+) –the probability of true positive predictions against false positive predictions, indicates the ability for a model to rule in disease[202].

When dealing with highly imbalanced data, reporting measures of accuracy can be misleading (discussed in Chapter 1.3.5). In such cases, measures of balanced accuracy (the average accuracy for each outcome class), F_1-score (the harmonic mean of precision and recall), or prevalence insensitive measures such as sensitivity, specificity and AUC are more appropriate for evaluating the performance of a model[134]. Hence, all performance metrics detailed above were reported when evaluating the prediction models developed in this thesis.

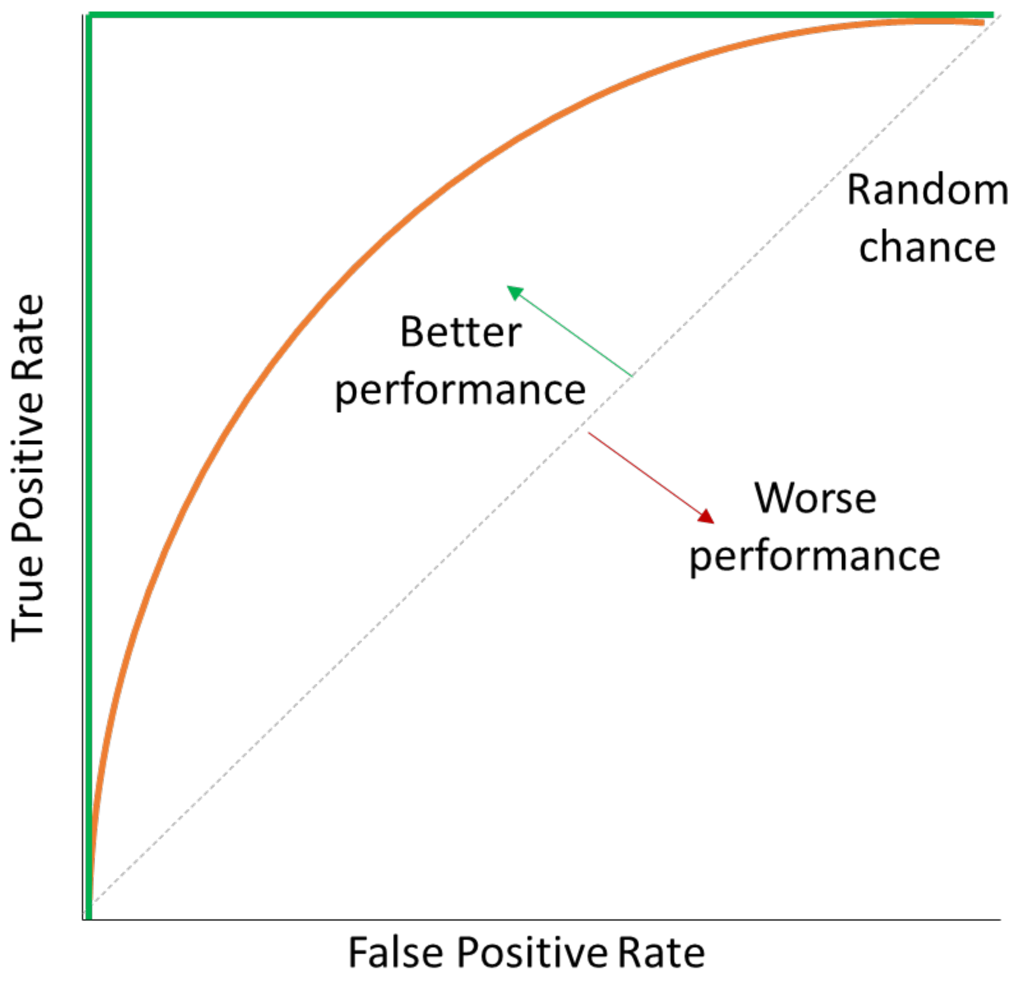


Figure 2.7 Schematic of a Receiver Operating Characteristic curve

The area under the ROC curve (AUC) is a plot of the true positive rate against the false positive rate across all classification threshold cut-offs. The dashed grey line indicates a model that is no better than chance (AUC=0.5). A model with good performance (orange line) has an AUC greater than 0.5, with better models having curves towards the top-left corner. The solid green line represents a perfect classification model (AUC=1.0). Poor models will have curves towards the bottom-right corner of the plot. Figure reproduced based on Hajian-Tilaki[204] (copyright license: CC-BY-NC 4.0).

2.3.7 Interpreting “black box” machine learning models

As previously described, a number of methods have recently been proposed to address the poor interpretability of “black-box” machine learning models that significantly hinder their application in healthcare[167,168,206]. Shapley Additive exPlanations (SHAP) is a tool, which unifies a number of model interpretability methods, capable of explaining the decisions of any machine learning model[168].

SHAP is often considered an extension of Shapley values, a concept rooted in coalitional game theory, which aims to determine the average contribution of each feature in offering a certain prediction. Shapley values are defined as the average marginal contribution of a feature value across all possible coalitions (all possible combinations of features in the model). The average contribution of each predictor is approximated based on the difference in the predictions obtained from the inclusion and exclusion (random assignment) of the predictor, averaged across all possible coalitions of the model. As a result, the Shapley value for each predictor is not the difference in the prediction if that feature was removed. Rather, it is the contribution of the feature to the prediction of a particular instance compared to the average prediction for the dataset. SHAP computes the Shapley values for each predictor of the original model and represents them in a linear model, as an additive feature attribution method[168,206].

In this thesis, SHAP was used for three purposes: i) to infer feature importance and effect (direction of risk for developing asthma) for the subset of predictors identified from the feature selection process; ii) to gain insight into the global explanation of each prediction model; and iii) to obtain local explanations of individual predictions.

2.4 Software

For this thesis, all data cleaning and encoding was performed using R statistical programming language (version 3.5.1)[207]. Quality control of the IOWBC genotype data was performed using Bash script and PLINK (version 1.90)[208,209]. DNA methylation data in the IOWBC was pre-processed using R (version 3.6.1) and specific packages as detailed in Chapter 2.1.1.3.

The development of asthma prediction models using machine learning approaches (detailed in Chapter 5 and Chapter 6), was primarily conducted using Python scripting language (version 3.6.8). The scikit-learn[127] and boruta[193] packages were used to perform feature selection using the RFE and Boruta methods, respectively. The balanced random forest algorithm used for feature selection was sourced from the imbalanced-learn package[189]. All machine learning algorithms were developed and tuned using the scikit-learn package[127], except the naïve Bayes algorithm which used the mixed-naïve-Bayes package[210]. Oversampling by ADASYN was also performed using the imbalanced-learn package[189]. SHAP values were computed for the random forest algorithm used during RFE as well as to explain individual predictions made by the final asthma prediction models using the SHAP packages TreeExplainer[211] and KernelExplainer[168,212], respectively. The statistical programming language R (version 3.5.1)[207] was used to implement MICE imputation (mice package[201]).

R statistical programming language was used for the remaining analyses performed in this thesis[207]. This included all analyses for the validation of existing models detailed in Chapter 4 (R version 3.5.1) and to construct the polygenic and methylation risk scores for childhood asthma detailed in Chapter 6 (R version 3.6.1).

In addition, the IRIDIS High Performance Computing Facility and associated support services at the University of Southampton were utilised to undertake this work.

Chapter 3 Systematic Review of Existing Childhood Asthma Prediction Models

3.1 Introduction

Asthma symptoms usually manifest in early life. However, with no clear disease trajectory or definitive test to confirm an asthma diagnosis before the age of 5 (as previously discussed in Chapter 1), it is difficult to predict which preschool children will develop asthma later in childhood and which children will see their symptoms subside. Unsurprisingly, there is a window of uncertainty in clinical decision-making[40], resulting in both under- and over-diagnosis of probable asthmatic preschoolers[112,113]. Prediction models which can identify true future asthmatics from a group of high-risk, symptomatic preschool children can assist physicians in providing early diagnoses and interventions. However, models which can also identify future asthmatics within a general population of preschoolers have the additional benefits of identifying late-onset asthmatics and stratifying individuals by asthma risk to subsequently promote asthma prevention among moderate/low-risk children. Besides being cost-effective, such strategies, as already demonstrated in other disease areas[93,99-101], could promote personalised asthma care, limit unnecessary exposure to the adverse effects of asthma medications, and reduce the wastage of healthcare resources[112].

In addition, to be of clinical value, the performance of any predictive model needs to be reproducible in independent populations with comparable performance characteristics. Although several prediction models for childhood asthma exist, not all have been validated in independent populations. Surprisingly, none have yet been incorporated into clinical practice[213-215].

3.1.1 Objectives

In this chapter, a systematic review has been conducted to critically evaluate existing prediction models for school-age asthma development by assessing their predictive performance, statistical methodology and potential clinical utility (addressing Aim 1 of this thesis, detailed in Chapter 1.4). Where relevant, external validation studies of these models are also assessed. Finally, potential issues, which might be responsible for the lack of clinical utility of existing asthma prediction models, are identified and recommendations for future research priorities presented.

3.2 Methods

This systematic review (PROSPERO registration number: CRD42019146638) was conducted in accordance with the guidelines reported in the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement[216].

3.2.1 Search strategy

An electronic search of three databases: Medline, Embase, and Web of Science Core Collection was performed on 26th July 2019. Free-text and MeSH terms were used to identify articles related to predictive modelling for childhood asthma (Table B1-3).

All articles underwent a two-stage duplicate removal: first electronically using EndNote X8.2[217] followed by a manual removal of remaining duplicates. Two independent reviewers conducted a title and abstract screening to assess the relevance of the remaining articles. Discrepancies were resolved through discussion among the reviewers. A full-text and additional screening of citations in selected papers and reviews of prediction models for childhood asthma were conducted. Identified studies underwent data extraction and qualitative analysis.

3.2.2 Study selection

Articles were included if they met the following criteria: the study detailed the development of a novel prediction model or updated a pre-existing model; the target population was children aged ≤5 years; the main prediction outcome was future childhood asthma or wheeze persistence at school-age (6-13 years old); and at least two risk predictors were used to construct the model. Models developed in both general and high-risk populations were considered. Validation studies which improved upon existing models were included. Studies which externally validated existing models in populations unrelated to that in which they were developed were also included.

Articles were excluded if a final prediction score was not developed or studies failed to report any performance measures for model evaluation. Conference papers, randomised control trials, letters, editorials and non-English articles were excluded.

3.2.3 Data extraction

Information on study design, candidate predictors, statistical methodology for model development and the prediction outcome were collected from model derivation studies.

Model performance was evaluated using prediction measures of: discrimination, sensitivity, specificity, positive and negative predictive values (PPV and NPV, respectively) and positive and negative likelihood ratios (LR+ and LR-, respectively) (described in Chapter 2.3.6). Where absent, likelihood ratios were calculated using reported sensitivity and specificity. Where applicable, performance measures were collected from both derivation and validation studies in order to assess model generalisability. The Prediction model Risk Of Bias ASsessment Tool (PROBAST) checklist[218] was used to critically appraise the risk of bias and applicability of each article.

3.3 Results

The literature search identified 4187 articles (Figure 3.1). Following the removal of 1204 duplicate articles, 2983 articles underwent title and abstract screening. The screening process identified 59 articles for full-text review. Of these, 25 studies were deemed relevant. An additional citation screening of relevant articles and the seven identified review articles on childhood asthma prediction models identified a further three studies. These 28 studies were classified into two categories based on the methods used for developing the prediction models: logistic regression-based (n=20) and machine learning approaches (n=4). The remaining four studies were external validations of previously developed models.

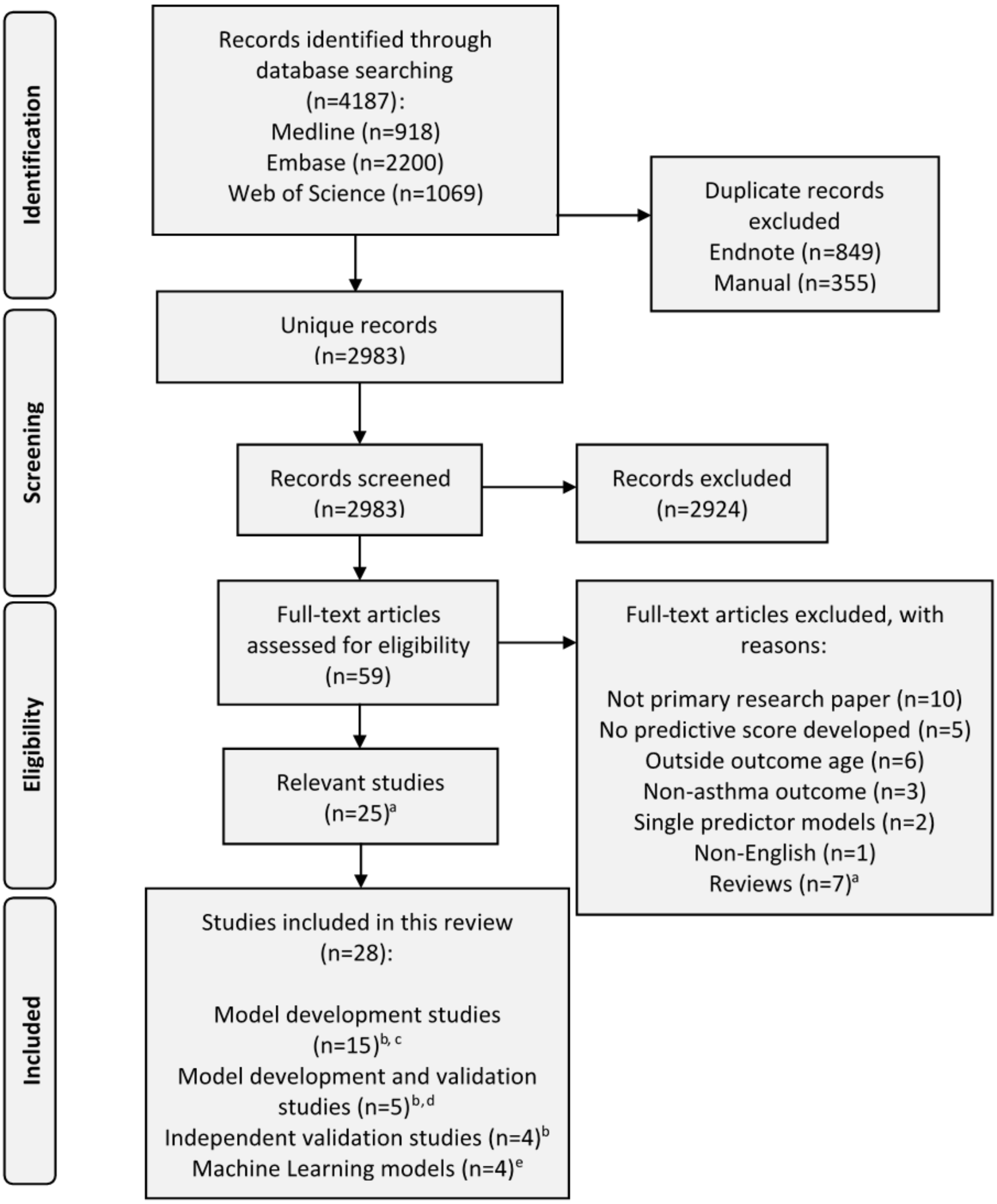

Identification
Screening
Eligibility
Included
Records identified through database searching (n=4187): Medline (n=918) Embase (n=2200) Web of Science (n=1069)
Duplicate records excluded Endnote (n=849) Manual (n=355)
Unique records (n=2983)
Records screened (n=2983)
Records excluded (n=2924)
Full-text articles assessed for eligibility (n=59)
Full-text articles excluded, with reasons:
Not primary research paper (n=10)
No predictive score developed (n=5)
Outside outcome age (n=6)
Non-asthma outcome (n=3)
Single predictor models (n=2)
Non-English (n=1)
Reviews (n=7)a
Relevant studies (n=25)a
Studies included in this review (n=28):
Model development studies (n=15)b, c
Model development and validation studies (n=5)b,d
Independent validation studies (n=4)b
Machine Learning models (n=4)e

Figure 3.1 PRISMA flow diagram of the systematic review search strategy

[a] Citation screening of articles and existing reviews identified three additional studies

[b] Included in the final qualitative analysis

[c] One study transformed a diagnostic model into a prediction model upon external validation (considered a developmental study in this review).

[d] Validated the developmental study model (n=2) or an existing model (n=3).

[e] Excluded from the main qualitative analysis

3.3.1 Regression-based models

Twenty-one regression-based, specifically logistic regression, prediction models were described in 20 studies (Figure 3.1). Thirteen of the 21 models were novel whilst eight were modifications of existing models: six modified the Asthma Predictive Index (API)[219-224]; one updated the PIAMA risk score[225] and one adapted the Obstructive Airway Disease (OAD) risk score[226]. Additionally, nine studies externally validated six prediction models, detailed within either developmental (n=5) or independent validation studies (n=4)[227-230].

Table 3.1 Summary of existing childhood asthma prediction models developed using regression-based approaches

Risk score	Year	Target population, age	Prediction population, age	Study size, prevalence (n,%)[a]	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR-	Discrimination
Independently validated prediction models											
Loose Asthma Predictive Index (API)[c] [219]	2000	General population, ≤3	6, 8, 11, 13	986 (57.1)	41.60	84.70	59.10	73.20	2.72[h]	0.69[h]	-
Stringent Asthma Predictive Index (API)[b,c] [219]	2000	General population, ≤3	6, 8, 11, 13	1002 (57.1)	15.70	97.40	76.60	68.30	6.04[h]	0.87[h]	-
Prevalence and Incidence of Asthma and Mite Allergy (PIAMA)[d,f,i] [231]	2009	High-risk[l], 0-4	7-8	2171 (11.1)	19.00	97.00	42.00	91.00	6.33[h]	0.84[h]	0.72
Persistent Asthma Predictive Score (PAPS)[f] [232]	2011	High-risk[l], <2	6	200 (47.5)	42.40	89.60	66.70	75.90	4.06	0.64	0.66
Predicting Asthma Risk in Children (PARC) Tool[d,f,i] [233]	2014	High-risk[l], 1-3	6-8	1226 (28.1)	72.00	71.00	49.00	86.00	2.47	0.40	0.74
Paediatric Asthma Risk Score (PARS)[f,i] [234]	2018	High-risk[j], ≤3	7	589 (16.1)	68.00	77.00	37.00	93.00	3.02	0.41	0.80

Risk score	Year	Target population, age	Prediction population, age	Study size, prevalence (n,%)[a]	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR-	Discrimination
Exploratory prediction model studies											
Modified Asthma Predictive Index (mAPI)[b,c,e] [223,224]	2004	High risk[j], 2-3	4-6	259 (28.2)	17.00	99.00	-	-	21.00	0.84	-
Singer *et al*. risk score (API + FeNO)[b,d] [221]	2013	High riskk, ≤ 4	6-10	166 (41.0)	75.00	62.30	58.00	78.20	1.99[h]	0.40[h]	-
Modified Asthma Predictive Index (m^2API)[b,c] [223]	2014	High risk[j], 1-3	6, 8, 11	259 (28.2)	30.00	98.00	-	-	16.00	0.71	-
University of Cincinnati (ucAPI)[b] [220]	2014	High risk[j], 3	7	589 (17.5)	44.00	94.10	60.30	89.30	7.50	0.60	-
Klaassen *et al*. (API + biomarkers)[b,d,i] [222]	2015	High risk[l], 2-4	6	198 (38.4)	88.00	90.00	90.00	89.00	8.80[h]	0.13[h]	0.86
Recurrent Wheeze Score (Isle of Wight, IoW) [235]	2003	High risk[l], 4	10	1034 (12.1)	52.50	84.60	68.40	73.70	3.41[h]	0.56[h]	-
Eysink *et al*.(RAST)[d] [236]	2005	High risk[k], 1-4	6	123 (26.8)	-	-	-	-	-	-	0.87
Obstructive Airway Disease (OAD) score for asthma[h] [237]	2007	General population [m], 2	10	449 (21.6)	55.60	86.40	52.90	87.60	4.09[h]	0.51[h]	0.78
Combined IgE antibodies and OAD (OAD + IgE)[b] [226]	2010	General population [m], 2	10	371 (50.0)	-	-	-	-	-	-	-
Updated PIAMA[b,f] [225]	2013	High-risk[k], 0-4	6-7	5048 (5.5)	63.80	73.90	12.40	97.20	2.44	0.49	0.75

Risk score	Year	Target population, age	Prediction population, age	Study size, prevalence (n,%)[a]	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR-	Discrimination
Lødrup Carlsen *et al.* [238]	2010	General population, at birth	10	607 (11.0)	64.00	67.00	19.00	94.00	1.94[h]	0.54[h]	0.72
Clinical Asthma Prediction Score (CAPS)[d,f,i] [239]	2013	High-risk[l], 1-5	6	438 (42.7)	-	-	74.30	78.40	-	-	0.73
Boersma *et al.* [240]	2017	High risk[l], 1-3	6	116 (62.9)	-	-	-	-	-	-	0.79
Szentpetery *et al.*[g] [241]	2017	General population, 1-4	8	2339 (5.0)	-	-	-	-	-	-	-
MAAS Asthma Prediction Tool (MAAS-APT)[d,f,i] [242]	2018	High-risk[l], 3	8-11	336 (34.8)	47.00	93.00	75.00	78.00	6.30	0.60	0.79

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; MAAS: Manchester Asthma and Allergy Study, RAST (radio-allergosorbent test).

[a] Prevalence = proportion of cases with the outcome included in the selected study sample – for the IoW model which details a stratified outcome of wheeze, prevalence of persistent wheeze is reported; prevalence for the loose and stringent API refers to the reported active asthma in at least one survey within the prediction window.

[b] Prediction models which modified a previous model.

[c] The loose and stringent API, mAPI and m^2API were evaluated at 6, 8, 11 and 13 (loose and stringent API only) years. Study details and performance measures are given for asthma prediction in at least one survey within the prediction window for the loose and stringent API and at age 6 for the

mAPI/m^{2}API.

[d] Internal validation during model development was performed using bootstrapping (API+biomarkers, API+FeNO, MAAS-APT, RAST, PIAMA and CAPS) or leave-one-out cross validation (PAPS). Where applicable, the internal validation performance measures are reported. Unbootstrapped discrimination was reported for the API+biomarkers (AUC=0.95), RAST (AUC=0.87), and PIAMA risk score (AUC=0.74).

[e] Performance measures extracted from Chang *et al*.'s study.

[f] Models provided performance measures over a range of thresholds. Performance measures are reported at the threshold recommended in their developmental study.

[g] The study initially developed a diagnostic model targeting and predicting childhood asthma at age 6. For external validation in the BAMSE cohort, this model was transformed into a prediction model targeting children between ages 1-4 to predict asthma at age 8. The latter model was considered as a developmental study in this review and study details are reported for the BAMSE population in which the prediction model was evaluated.

[h] Where unreported, likelihood ratios were calculated based on reported sensitivity and specificity as: LR+ = sensitivity/ 1- specificity, LR- = 1- sensitivity/ specificity.

[i] Model calibration was evaluated in the study.

High-risk study cohort specified by parental history of allergy/asthma ([j]), presence of asthma-like symptoms ([k]), presence of asthma-like symptoms, specifically wheeze ([l]).

[m] Nested case-control study within a general population birth cohort of children age 2 with obstructive airway disease.

3.3.1.1 Target population

Of the 21 models carried forward for qualitative analysis (Table 3.1), six were developed in the general population[219,226,237,238,241] and 15 within high-risk populations, the latter restricting inclusion to children with a parental history of allergy/asthma (four models)[220,223,224,234] or asthma-like symptoms (11 models[225,236], with nine specifically targeting children experiencing wheeze[221,222,231-233,235,239,240,242]). Only one model was derived based on predictors initially associated with childhood asthma within a low-income, Puerto-Rican population[241].

3.3.1.2 Predictors

Thirty-eight different predictors were used among the 21 identified models, including seven variations of wheeze and two different measures for both allergic sensitisation and pulmonary function (Table 3.2). The number of predictors used to construct the models ranged between three and ten. Twenty out of 38 predictors were each included in just one of the 21 models (last column, Table 3.2). For example, familial pollen allergy was a predictor in Eysink *et al*.'s model alone, while race was only included in the PARS model. A history of parental asthma and personal eczema were the most frequently used predictors of childhood asthma, each incorporated into 14 models. Three studies used data only available in early life (≤2 years)[226,232,237] whilst another only used predictor data collected at birth[238]. Predictor information was mainly collected from parent-reported questionnaires or standard clinical assessments. Sixteen models required data from additional clinical tests such as blood or skin prick tests (SPT) to assess allergic sensitisation status (14 models); measures of pulmonary function (two models); biomarkers of volatile organic compounds (VOCs) in exhaled breath condensate (one model); and gene expression in peripheral blood (one model).

Table 3.2 Summary of predictors included in the regression-based childhood asthma prediction models

	Loose API [219]	**Stringent API** [219]	**mAPI** [224]	**m2API** [223]	**API + FeNO** [221]	**API + biomarkers** [222]	**ucAPI** [220]	**IoW** [235]	**Eysink *et al*** [236]	**OAD** [237]	**OAD + IgE** [226]	**PIAMA** [231]	**Updated PIAMA** [225]	**Lødrup Carlsen *et al.*** [238]	**PAPS** [232]	**PARC** [233]	**CAPS** [239]	**Boersma *et al.*** [240]	**Szentpetery *et al.*** [241]	**MAAS-APT** [242]	**PARS** [234]	**Total**[a]
Subject characteristics																						
Sex												X	X	X		X			X			**5**
Age									X							X	X					**3**
Race																					X	**1**
Gestation length												X	X									**2**
Clinical Symptoms																						
Any wheeze									X													**1**
Early wheeze	X				X																X	**3**
Frequent wheeze			X	X			X					X	X			X						**6**
Early frequent wheeze		X																				**1**
Exercise-induced																X				X		**2**
Aeroallergen-induced																X						**1**
Wheeze without cold	X	X	X	X	X	X	X					X	X			X	X				X	**12**
Eczema	X	X	X	X	X	X	X					X	X		X	X		X		X	X	**14**

	Loose API [219]	**Stringent API** [219]	**mAPI** [224]	**m2API** [223]	**API + FeNO** [221]	**API + biomarkers** [222]	**ucAPI** [220]	**IoW** [235]	**Eysink *et al*** [236]	**OAD** [237]	**OAD + IgE** [226]	**PIAMA** [231]	**Updated PIAMA** [225]	**Lødrup Carlsen *et al.*** [238]	**PAPS** [232]	**PARC** [233]	**CAPS** [239]	**Boersma *et al.*** [240]	**Szentpetery *et al.*** [241]	**MAAS-APT** [242]	**PARS** [234]	**Total**[a]
Allergic rhinitis	X	X			X	X	X												X			**6**
Shortness of breath																X				X		**2**
Nasal symptoms								X														**1**
Nocturnal symptoms																	X					**1**
Cough on exertion																				X		**1**
Recurrent chest infection								X														**1**
Respiratory tract infection												X										**1**
Hospital admission for respiratory symptoms										X	X							X				**3**
Number of BO episodes										X	X											**2**
Duration of BO										X	X											**2**
Disturbances to activity																X						**1**
Obesity																			X			**1**

	Loose API [219]	Stringent API [219]	mAPI [224]	m2API [223]	API + FeNO [221]	API + biomarkers [222]	ucAPI [220]	IoW [235]	Eysink *et al* [236]	OAD [237]	OAD + IgE [226]	PIAMA [231]	Updated PIAMA [225]	Lødrup Carlsen *et al.* [238]	PAPS [232]	PARC [233]	CAPS [239]	Boersma *et al.* [240]	Szentpetery *et al.* [241]	MAAS-APT [242]	PARS [234]	Total[a]
Familial characteristics																						
Parental asthma	X	X	X	X	X	X	X	X					X		X	X	X		X		X	**14**
Familial allergy to pollen									X													**1**
Parental inhaled medication												X										**1**
Alcohol intake during pregnancy														X								**1**
Environmental exposures																						
Parental education												X	X									**2**
Ease of acquiring a babysitter														X								**1**
Family network														X								**1**

	Loose API [219]	**Stringent API** [219]	**mAPI** [224]	**m2API** [223]	**API + FeNO** [221]	**API + biomarkers** [222]	**ucAPI** [220]	**loW** [235]	**Eysink *et al*** [236]	**OAD** [237]	**OAD + IgE** [226]	**PIAMA** [231]	**Updated PIAMA** [225]	**Lødrup Carlsen *et al.*** [238]	**PAPS** [232]	**PARC** [233]	**CAPS** [239]	**Boersma *et al.*** [240]	**Szentpetery *et al.*** [241]	**MAAS-APT** [242]	**PARS** [234]	**Total**[a]
Clinical tests																						
Blood eosinophilia	X	X	X	X																		**4**
Specific IgE (RAST or other assay)			X	X		X			X		X				X		X	X				**8**
Skin prick test (SPT)							X	X												X	X	**4**
Fraction of exhaled nitric oxide (FeNO)					X																	**1**
Lung function (V_E)[c]														X								**1**
Exhaled volatile organic compounds						X																**1**
Gene expression						X																**1**
Total[b]	**6**	**6**	**6**	**6**	**6**	**7**	**6**	**4**	**4**	**3**	**4**	**8**	**7**	**5**	**3**	**10**	**5**	**3**	**6**	**5**	**6**	

[a] Total number of models that use each predictor

[b] Total number of predictors included in each model

[c] V_E = minute ventilation

BO = bronchial obstruction

3.3.1.3 Outcome

The prediction outcome in most studies (19/20) was school-age asthma, yet nine different definitions of asthma were used (Table 3.3). Seventeen studies included asthma-like symptoms, twelve included a doctor diagnosis and nine incorporated objective pulmonary tests as components in their asthma definition. One study used persistent wheeze, determined through the frequency of wheezing episodes, as the prediction outcome[235]. The most common definition (in 5/20 studies) specified a combination of asthma-like symptoms, use of asthma medications and/or objective respiratory tests. All studies identified a child's asthma status by evaluating the outcome criteria within the last 12 months except one which evaluated the asthma criteria across two consecutive years[231].

Table 3.3 Summary of the main asthma definitions used amongst the childhood asthma prediction model developmental studies

Asthma outcome definitions	Number of studies	Study reference
Doctor diagnosis only	1	[221]
Symptoms only	1	[235]
Doctor diagnosis and symptoms	4	[219,225,241,242]
Doctor diagnosis and medication	2[e]	[223]
Symptoms and medication	2	[233,240]
Doctor diagnosis, symptoms, medication	1	[231]
Symptoms, medication, lung function tests[a]	5	[222c,d], [220c,d], [236 d], [239c,d], [238b]
Doctor diagnosis, symptoms, lung function tests[a]	1	[234b,d]
Doctor diagnosis, symptoms, medication, lung function tests[a]	3	[237b], [226b], [232c]

[a] Lung function tests comprised of one or a combination of: exercise challenge tests ([b]), spirometry assessing reversibility to bronchodilators ([c]) and bronchial hyper-responsiveness to methacholine or histamine ([d]).

[e] The asthma outcome for the mAPI was extracted from the m^{2}API study which evaluated the model's performance.

3.3.1.4 Model construction

The API and its modifications are clinical indices requiring a combination of major and minor criteria to be met. The other prediction models are weighted scoring systems based on derivations of each predictor's regression coefficients, with the exception of two unweighted scoring systems[235,241].

3.3.1.5 Performance measures

Three studies failed to report any model performance measures detailed in Chapter 3.2.3. Of these, the modified Asthma Predictive Index (mAPI), developed within a randomised clinical trial protocol, did not evaluate the model's performance[224]. Performance measures for the mAPI were extracted from the study by Chang *et al*.[223] which evaluated and compared the mAPI against another modified API (m^2API)[223]. The other two studies only reported single performance measures of population attributable risk[241] and Nagelkerke R^2 [226].

Discriminative ability was reported for 12 models and ranged between 0.66 and 0.87. Sixteen models reported sensitivity (range: 15.7-88%) and specificity (range: 62.3-99%). PPV and NPV were reported for 15 models, ranging between 12.4-90% and 68.3-97.2%, respectively. Likelihood ratios were reported for eight models and derived for an additional eight models using reported sensitivity and specificity. The ability to rule in disease (LR+) ranged from 1.94-21 whilst the ability to rule out disease (LR-) ranged from 0.13-0.87.

3.3.1.6 Validation

Nine studies performed external validation: four validated the loose and/or stringent API, two validated the PIAMA and PARC models whilst the PAPS and PARS models were each validated once (Table 3.4). Upon validation, most models demonstrated a trade-off between improvements in sensitivity at the expense of specificity, resulting in increased false positive predictions and a decline in PPV and LR+ estimates compared to their derivation models. Whilst the PARS model showed comparable performance upon validation, only the PARC model demonstrated superior performance, with improvement in LR+ (2.47 vs 2.63) and AUC (0.74 vs 0.83) compared to its developmental model.

Table 3.4 Summary of the performance of the childhood asthma prediction models upon independent validation

Developmental model	Author	Population geography	Risk group	Variation in predictors	Variation in outcome	Study size	Study asthma prevalence (%)	Target age	Prediction age	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR-	Discrimination
Loose API	Castro-Rodriguez *et al.*[219]	USA	General population			986	57.1	≤3	6-13	41.6	84.7	59.1	73.2	2.72[a]	0.69[a]	-
	Rodriguez-Martinez *et al.*[227]	Colombia	High risk	-	-	93	22.5	1-3	5-6	71.4	33.3	23.8	80	1.07	0.86	-
	Leonardi *et al.*[230]	UK	General population	✓		1731	11.5	2-3	7	57	80	26	94	2.85[a]	0.54[a]	0.68
						1291	10.5	2-3	10	57	81	25	94	3.00[a]	0.53[a]	0.69
	Devulapalli *et al.*[237]	Norway	General population	✓	✓	459	21.1	3	10	59.8	79	43.9	87.7	2.85[a]	0.51[a]	-

Developmental model	Author	Population geography	Risk group	Variation in predictors	Variation in outcome	Study size	Study asthma prevalence (%)	Target age	Prediction age	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR-	Discrimination
Stringent API	Castro-Rodriguez *et al.*[219] [a]	USA	General population			1002	57.1	≤3	6-13	15.7	97.4	76.6	68.3	6.04[b]	0.87[b]	-
	Rodriguez-Martinez *et al.*[227]	Colombia	High risk	-	-	93	22.5	1-3	5-6	42.9	79.2	37.5	82.6	2.06	0.72	-
	Leonardi *et al.*[230]	UK	General population	✓	-	1683	11.5	2-3	7	37	93	40	93	5.29[b]	0.68[b]	0.65
						1257	10.5	2-3	10	32	94	35	92	5.33[b]	0.72[b]	0.63
	Caudri *et al.*[231]	Netherlands	High risk	✓	✓	1177	11.7	0-4	7-8	20	92	25	90	2.50[b]	0.87[b]	0.62
	Devulapalli *et al.*[237]	Norway	General population	✓	✓	459	21.1	3	10	56.7	83	47.8	87.4	3.34[b]	0.52[b]	-
PIAMA	Caudri *et al.*[231] [a]	Netherlands	High risk			2171	11.1	0.4	7-8	19	97	42	91	6.33[b]	0.84[b]	0.74
	Hafkamp-de Groen *et al.*[225]	Netherlands	High risk	✓	✓	2877	6.0	1-4	6	-	-	-	-	-	-	0.74
	Rodriguez-Martinez *et al.*[227]	Colombia	High risk	-	✓	123	53.6	1-3	5-6	54.5	78.9	75.0	60	2.59	0.58	-

Developmental model	Author	Population geography	Risk group	Variation in predictors	Variation in outcome	Study size	Study asthma prevalence (%)	Target age	Prediction age	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR-	Discrimination
PARC	Pescatore *et al.*[233] [a]	UK	High risk			1226	28.1	1-3	6-8	72	71	49	86	2.47	0.40	0.74
	Grabenhenrich *et al.*[228]	Germany	High risk	✓	-	140	20.0	3	8	82	69	40	94	2.63	0.26	0.83
	Pedersen *et al.*[229]	UK	High risk	✓	-	2690	14.0	1.5-3.5	7.5	69	76	32	94	2.87	0.41	0.77
PAPS	Dupuy et al.[232] [a]	France	High risk			200	47.5	<2	6	42.4	89.6	66.7	75.9	4.06	0.64	0.66
	Dupuy et al.[232]	France	High risk	-	-	227	18.9	<2	13	62.8	67.4	31	88.6	1.93[b]	0.55[b]	0.65
PARS	Biagini Myers *et al.*[234] [a]	USA	High risk			589	16.1	≤3	7	68	77	37	93	3.02	0.41	0.80
	Biagini Myers *et al.*[234]	UK	General population	✓	✓	1098	-	2	10	67	79	36	93	3.25	0.41	0.79

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio.

[a] Prediction models as reported in the developmental studies; unmarked rows refer to performance reported in the external validation studies.

[b] Likelihood ratios calculated based on reported sensitivity and specificity as: LR+ = sensitivity/ 1- specificity, LR- = 1- sensitivity/ specificity.

✓ Used altered definitions in the external validation study compared to the original developmental study – predictors= exclusions or surrogate variables used; outcome= variation in components used to determine asthma.

3.3.1.7 Critical appraisal

The overall risk of bias was deemed high for all 21 models due to: i) predictor and outcome bias (21 and 17 models respectively), predominantly due to the subjective interpretation of their definitions, particularly those based on parent-reported information; and ii) biased analysis due to an inappropriate number of candidate predictors, inappropriate handling of missing data, failure in reporting performance measures (e.g. calibration) or failure in treating models for potential overfitting or performance optimisation as detailed in the PROBAST checklist (Table 3.5). The 15 studies which used high-risk developmental populations presented with low risk of bias (assuming their intended use in settings similar to their developmental study) but high concern regarding applicability to a general population.

3.3.2 Machine learning approaches

Four studies which utilised machine learning approaches to develop five prediction models for childhood asthma within a single paediatric hospital population of diagnosed asthma patients were identified[163-166]. These studies presented with ambiguity in their study design with regards to unclear predictor definitions, time-points of predictor measurements and population characteristics. Additionally, due to limitations of using an asthma diagnosis as a predictor, the small study size for machine learning applications, and signs of overfitting in the reported results, these studies were excluded from the main qualitative analysis. However, they were included in the review to highlight novel methodologies currently being explored for childhood asthma prediction (Table 3.6).

Table 3.5 Critical appraisal for the bias and applicability of each study using the PROBAST checklist

	Loose API[219]	**Stringent API**[219]	**mAPI**[224]	**m^{2}API**[223]	**API + FeNO**[221]	**API + biomarkers**[222]	**ucAPI**[220]	**loW**[235]	**Eysink *et al.*** [236]	**OAD**[237]	**OAD + IgE**[226]	**PIAMA**[231]	**Updated PIAMA**[225]	**Lødrup Carlsen *et al.***[238]	**PAPS**[232]	**PARC**[233]	**CAPS**[239]	**Boersma *et al.***[240]	**Szentpetery *et al.***[241]	**MAAS-APT**[242]	**PARS**[234]
Risk of bias																					
Participants	L	L	H	H	H	H	H	H	H	L	L	H	H	L	H	H	H	H	L	H	H
Predictors	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
Outcome	H	H	H	H	H	L	H	H	H	L	L	H	H	H	U	H	H	H	H	H	H
Analysis	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
Overall	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**
Concern regarding applicability																					
Participants	L	L	H	H	H	H	H	H	H	L	L	H	H	L	H	H	H	H	L	H	H
Predictors	L	L	L	L	L	H	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
Outcome	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
Overall	**L**	**L**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**L**	**L**	**H**	**H**	**L**	**H**	**H**	**H**	**H**	**L**	**H**	**H**

Risk of bias and applicability were assessed as: H = high, L=low, U= unclear using the criteria outlined in the PROBAST checklist[218]. For each domain, the risk of bias or concern of applicability is considered: high - if ≥1 signalling question in the PROBAST criteria was answered “no” or “probably no”; low – if the answer to all signalling questions was “yes”; unclear – if relevant information was missing to answer the signalling question and none of the signalling questions were answered “no”. The overall risk of bias and applicability were deemed low if all domains were evaluated as low risk, high risk if ≥1 domain was considered high-risk, unclear if ≥1 domain was considered unclear and all other domains were low-risk.

Table 3.6 Summary of the identified childhood asthma prediction models developed using machine learning approaches

Study	Feature selection	Number of predictors	Study size	Model Algorithm	Accuracy	Sensitivity	Specificity	PPV	NPV	LR+	LR-	AUC	External validation
[163]	Correlation analysis	10	112	Multilayer Perceptron	1.00	1.00	1.00	-	-	NA[a]	0.00[a]	-	No
[164]	Genetic algorithm	4	112	Artificial Neural Network	0.95	-	-	-	-	-	-	-	No
[165]	Principal Component Analysis	18	112	Least-square Support Vector Machine	0.96	0.95	0.96	-	-	21.64[a]	0.05[a]	-	No
[166]	Partial least square regression	9	112	Multilayer Perceptron	0.97	0.96	1.00	-	-	NA[a]	0.04[a]	-	No
[166]	Partial least square regression	9	112	Probabilistic Neural Network	0.97	1.00	0.80			5.00[a]	0.00[a]	-	No

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

[a] Likelihood ratios calculated based on reported sensitivity and specificity as: LR+ = sensitivity/ 1- specificity, LR- = 1- sensitivity/ specificity, NA= undefined

- Not reported

3.4 Discussion

This review identified 26 models for predicting childhood asthma at school-age, but none have been widely implemented into standard clinical practice. Only the API is mentioned in asthma management guidelines[2] and has been utilised with caution (upon modification), in the recruitment of participants into clinical trials[224]. Against this background, a critical evaluation of these studies aimed to identify potential problems surrounding the lack of applicability of these models. The key issues centred on: the choice of population for model derivation and/or validation, predictor and outcome definitions, methodologies employed for predictor selection, methods of data collection, study power, and the interpretability of models.

3.4.1 Choice of population

The performance of any predictive model is highly dependent on its developmental setting and may not generalise well in alternative risk populations. Fifteen of the twenty-one regression-based models were developed in high-risk populations. High-risk populations, which have a higher asthma prevalence compared to the general population, are commonly used for model development in the hope of increasing the power for predictor selection and the detection of true asthmatics. However, such models may overestimate asthma risk within the general population. At present, only PARS has assessed this and was able to show comparable predictive performance in high risk and general populations (with the general population validation of PARS utilising the IOWBC used in this thesis). In contrast, the loose and stringent API, developed in a general population, demonstrated a substantial improvement in sensitivity, although at the cost of increasing false positive predictions, when validated in high-risk populations (Table 3.4).

3.4.2 Population-specific predictors

Most models were developed in European/predominantly Caucasian cohorts. Exposures specific to less developed countries, such as poverty and pollution, are typically not considered as important predictors of asthma in these models due to inadequate representation of such populations within the study cohorts[243]. For example, Szentpetery *et al*. initially developed a diagnostic model, identifying gun violence and an unhealthy diet as predictors of childhood asthma in a Puerto Rican population. However, when validated as a prediction model in a Swedish cohort, data for these two predictors were unavailable, potentially due to low concern for these risk factors in this population, and were excluded from the model[241].

3.4.3 Prediction window

Due to the transient nature of asthma-like symptoms in early life, the evaluation of clinical predictors from 4-5 years of age is more predictive of school-age asthma[38]. However, for prediction models developed with the intention of preventing asthma development rather than targeting children for early therapeutic intervention, predictions made at 4-5 years may already be too late. Four models used predictor data available before age 2[226,237,238], but only one was externally validated[232]. The model developed by Lødrup Carlsen *et al*. only used predictor data collected at birth, however the need to perform neonatal lung function tests (rarely conducted outside of a research setting) greatly impairs its potential clinical applicability[238].

3.4.4 Data collection

Most studies collected predictor information through parent-completed questionnaires, a method prone to recall bias and misclassifications. For example, studies have shown that less than 60% of parents are able to correctly identify wheeze[244], with around one third of parents changing their answer after watching a recording of wheeze[245]. Such under/overestimations of parent-reported predictors can result in poor model performance compared to models that use data collected from physicians, healthcare records or objective measurements.

3.4.5 Predictor availability

Thirty-eight different predictors indicative of well-documented asthma risk factors were used across the 21 regression-based models. This variation reflects the inherent heterogeneity of childhood asthma across different populations and variability in predictor availability between studies. Sixteen models required additional clinical tests, most commonly blood and SPTs to determine a child's atopic status. These tests were the main amendment in four of the seven modified prediction models. Four other studies demonstrated that the addition of IgE as a predictor in their models improved predictive power compared to their models without IgE[226,236,239,240]. Whilst the predictive value of information on allergic sensitisation is well established in the literature[246], it is important to note that SPT are not routinely performed in children not suspected to be atopic, particularly in early life, and blood testing to measure levels of eosinophilia or serum IgE are invasive tests which require the use of additional healthcare resources. One modification of the API included biomarkers of VOCs in exhaled breath condensate and gene expression[222]; despite ranking second in terms of AUC (AUC=0.86, unbootstrapped AUC=0.95), the use of this model is unlikely to be feasible outside of a specialist/research setting. Models developed with predictors which are not readily available, or which require the use of additional healthcare resources, can be limited in their generalisability and potential clinical implementation.

3.4.6 Predictor selection

Methodology for the selection of predictors varied between the 20 regression-based studies. Models used either a priori knowledge[219,223,224], univariate analysis[219], multivariate regression analysis[221,225] or a combination of univariate and multivariate regression[220,231,235,240,242]. Despite the latter two-stage combination approach being an established method used across biomedical research, this method can introduce significant bias to the feature selection process due to inconsistencies between univariate and subsequent multivariate analyses[247,248]. To address this, some studies adopted a stepwise backward or forward selection multivariate regression approach[222,234,236,238,239,241], and the PARC model[233] utilised LASSO (Least Absolute Shrinkage and Selection Operator) - a regularisation method which shrinks the effect size of less important predictors to zero, thus only selecting a subset of the most predictive features[131]. However, none of these studies fully address the issue of multicollinearity between candidate predictors that can introduce noise and subsequently reduce model performance. Among the four machine learning studies identified, supervised and unsupervised machine learning algorithms were used for feature selection[163-166]. Indeed, machine learning algorithms, particularly those such as random forest, recursive feature elimination and genetic algorithms, are more robust in handling the relatedness between predictors and may promote better predictor selection compared to regression-based methods[131,132].

3.4.7 Outcome

Childhood asthma is often considered an umbrella term describing a syndrome of different respiratory symptoms[1]. In a study reviewing 122 childhood asthma cohort studies, van Wonderen *et al*. identified 60 different outcome definitions of childhood asthma. Differences in definitions resulted in only 61% agreement in the classification of individuals as asthmatic/non-asthmatic and large variation in asthma prevalence between 15-51%[249]. This was reflected in this current review which identified the use of nine different asthma outcome definitions across the 20 regression-based studies. With models developed to predict childhood asthma already predicting a subjective entity, the use of different asthma definitions only amplifies the variation of the classification outcome. This may have led to an artificial variation in the prevalence of asthma across studies influencing the construction, optimisation and subsequent performance of the predictive models.

Variation in outcome definitions also pose a problem for evaluating the generalisability of a model and comparing the performance between models. For example, Rodriguez-Martinez *et al*. conducted a comparative validation of the API and PIAMA risk score using outcome definitions as described in their respective developmental studies. They showed that, because children are frequently prescribed asthma medication without the doctor always explicitly confirming a diagnosis of 'asthma', the true asthma outcome for some patients varied between the two model validations - some patients classified as non-asthmatics for analysis with the API were reclassified as asthmatic for the analysis with the PIAMA risk score[227]. Hence, a consensus on an objective definition acceptable to the clinical and research community is essential.

3.4.8 Study power

Upon critical appraisal, at least eight studies were identified as lacking sufficient power to develop stable prediction models; these studies had a ratio of candidate predictors to total number of cases lower than recommended (at least 20 cases per candidate predictor) to achieve sufficient power[225,231,233,235,238,239,242]. Underpowered studies risk important predictors not being selected (under-fitting –Type II error), the incorrect selection of predictors (overfitting –Type I error) as well as the misrepresentation of the associated directionality between predictors and the outcome[205].

Compared to traditional regression methods, machine learning approaches possess superior power and resolution for pattern recognition. By allowing a larger number of candidate predictors to be considered and being more robust to the relatedness between predictors, there is potential to identify novel predictors and exclude redundant predictors which may have been previously overlooked by traditional feature selection approaches [117,118,123]. Despite the potential benefits offered by machine learning methods, the four machine learning studies identified in this review remain underpowered[163-166]. Further studies are necessary to determine whether machine learning approaches can develop better performing childhood asthma prediction models over regression-based methods.

3.4.9 Validation

Models tend to perform best within their developmental population. External validation studies, which assess the true performance of models in independent populations, are essential to assess the generalisability of a model. However, only six of the 21 identified regression-based models were externally validated. None of the five machine learning models described in the four identified studies were externally validated (Table 3.6). Whilst the PARS and PARC models demonstrated comparable performance when validated, the other models demonstrated poorer predictive performance, particularly in terms of PPV and likelihood ratios (Table 3.4). This may be due to inconsistencies between the derivation and validation study designs, mainly with regards to the predictor/outcome definitions and the exclusion or use of surrogate variables for unavailable predictor information. Validation of all existing models within a single independent population using a single outcome definition is necessary to standardise inconsistencies in study design and population effect to facilitate a comparative analysis between models. However, this remains difficult in practice due to the need for a reference population of sufficient size with data available for all 38 predictors.

3.4.10 Interpretability

At present, a quantitative evaluation of the performance of existing models is difficult as not all studies report the standard performance measures. Discrimination (AUC) is often used to compare the overall performance between models, with a discriminative threshold of 0.80 considered to identify a very good predictive model[202]. Three developmental models reached this threshold but only one, PARS, was externally validated. The good generalisability of PARS (AUC=0.79) has facilitated its transformation into an online interactive tool and mobile app for use by both physicians and parents[234].

However, using discrimination alone to compare model performance is inappropriate as models with similar AUC can show large variations in sensitivity and specificity. There is a clear trade-off between optimising both of these performance measures, with no one model able to achieve both high sensitivity and specificity. Therefore, clear aims of whether a model intends to optimise towards higher sensitivity or specificity for the future application of prevention or asthma symptom management, respectively, would benefit the evaluation of a model's predictive power and viability[202].

Finally, the API and its modifications provide a dichotomous outcome of asthma risk, whilst the remaining regression-based models present asthma risk across a range of potential scores, often stratifying individuals into groups of low, medium or high risk. However, physicians are already able to make similar predictions upon clinical assessment which may explain the lack of clinical uptake of existing models. The exploration of novel approaches, such as machine learning, for the development of prediction models with greater probabilistic resolution of an individual's asthma risk is warranted.

Yet, existing prediction models are not redundant – the use of well-performing, externally validated models could be considered for use in clinical trials to support the stratification of participants for inclusion or treatment allocation. These models are likely to offer superior predictions compared to trials currently utilising the API[224] or, more frequently, parental history, to assess asthma risk.

3.5 Conclusions and future recommendations

Based on the findings of this review, a number of key considerations are needed for the development of future prediction models.

1. Study design and data availability

Improving model generalisability across all population settings could be achieved by standardising predictor and outcome definitions across settings and addressing issues of population bias and data availability. Whilst the perfect solution would be to establish a single, general population, prospective cohort of sufficient size for model development alongside an independent reference population for validation, this is unrealistic.

Instead, studies should specify and closely match the developmental population of the model for its future application. Data should be collected using objective measurements and high-quality, standardised questionnaires with unambiguous descriptions which are consistent across both clinical and research settings. Where parent-reported data is used, clinical jargon should be deconstructed and/or be supported by auditory or visual aids to minimise recall bias and misclassification wherever possible.

In addition, only easily derivable and commonly available clinical predictors should be used. Whilst biomarkers can have high predictive power, their predictive benefit needs to be measured against the cost of test availability across different healthcare settings, patient/physician time and demand on healthcare resources. Yet, the exploration and identification of novel biomarkers, particularly in early life, may encourage the transition from asthma management to prevention.

2. Isolating predictors for model development

Due to the heterogeneity of childhood asthma, a number of candidate predictors have been associated with childhood asthma. One approach to identifying predictors for model development is to isolate a subset of the most frequently used predictors from previous studies. For example, parental asthma, eczema, wheeze without cold, specific IgE, frequent wheeze, allergic rhinitis and gender have been used in at least a quarter of existing models (Table 3.2). However, as previously discussed, population-specific influences and predictor selection methodological limitations exist in these studies. A better approach would be for future studies to utilise a robust predictor selection method (such as recursive feature elimination), which is sufficiently powered and able to address the multicollinearity between predictors, in order to distinguish strong predictors from redundant variables within their specific population.

3. Model development methodologies

The majority of existing studies have utilised regression-based methods and have developed a number of similar prediction models, few generalising well in independent populations, and none widely implemented into clinical practice. Alternative methods such as machine learning approaches have advantages over these statistical methods as already discussed, particularly with regards to addressing frequently overlooked concerns of predictor relatedness, distinguishing between predictive and redundant predictors, and improving the resolution of predictions. Such methods had not been adequately implemented at the time of this systematic review, hence future studies using robust study designs are needed to assess their potential benefits for childhood asthma prediction. Subsequent to this review, a number of studies applying machine learning approaches for the prognostic prediction of childhood asthma have been identified[250,251]. Whilst these studies often reported high predictive accuracy, overfitting appeared to be a common problem, with many studies subject to small sample sizes and none of the studies externally validating their machine learning models to assess model generalisability (discussed further in Chapter 5).

It is crucial for any developed model to undergo external validation within a population similar to its future application. Unvalidated models are not clinically useful and are largely limited as exploratory studies. Reporting of all standard performance measures in both the development and validation populations are necessary to evaluate a model's generalisability and subsequently promote its clinical application for predicting school-age asthma.

Chapter 4 External Validation of Existing Prediction Models in IOWBC

4.1 Introduction

Twenty-six prediction models for childhood asthma were identified by the systematic review conducted in Chapter 3. The five machine learning models identified at the time of the review were regarded as exploratory studies due to limitations in study design and external validation. Of the remaining regression-based models, only six have undergone validation in independent cohorts. It is known that performance measures reported in the developmental population for prognostic models are often overly optimistic. Hence, the validation of any developed prediction model in independent populations is essential to provide insight into the model's true predictive power and generalisability[252,253].

Whilst external validation in any independent population is encouraged, the generalisability of a model can vary depending on the independent population chosen. This is largely due to data availability and differences in data collection time-points and definitions of both predictors and the outcome. Performance is also influenced by the underlying characteristics of the population. Furthermore, there is often a lack of commonality between studies in terms of the populations used for development and/or validation (when conducted) as well as an inconsistency in reported performance measures. As the populations chosen for validation often differ, it is difficult to conduct accurate comparisons between models. Therefore, external validation of all existing models within a single independent population is necessary to standardise potential variations in study design and population effects arising from the use of different study populations. Validation in a single independent cohort would promote direct and accurate comparisons between models.

4.1.1 Objectives

To address the second aim of this thesis (detailed in Chapter 1.4), all existing prediction models for childhood asthma identified by the systematic review (and for which data was available) were implemented and compared within the single population of the IOWBC.

4.2 Methods

All 21 of the existing regression-based models were considered for validation in the IOWBC (detailed in Chapter 2.1.1). Models were selected for validation if information for all predictors included in the model were available in the IOWBC (Table 3.2). To maximise the models considered for validation, surrogate variables were used, where deemed appropriate based on expert clinical opinion and statistical testing to identify strong associations between predictors (detailed in Chapter 2.1.1.1).

For each model selected for validation, only individuals with complete predictor and outcome data were included in the analysis. For each individual, the risk score of each model was calculated and the prediction compared against his or her reported asthma diagnosis at age 10 (defined in Chapter 2.2).

Performance accuracy and measures of sensitivity, specificity, positive and negative predictive values, likelihood ratios, F_1-score and AUC were calculated based on the optimal cut-off threshold specified in the original model publications.

4.3 Results

Of the 21 regression-based prediction models developed for childhood asthma, five models had data available for implementation in the IOWBC - ucAPI[220], uPIAMA[225], Szentpetery *et al.*'s risk score[241], PARS[234] and IoW[235]. The implementation of the first four models can be considered to be external validation analyses. The IoW model was initially developed in the IOWBC to predict an outcome of persistent wheeze; therefore, although its inclusion in this analysis cannot be considered as an independent validation, it was included to assess the model's ability to predict the outcome of childhood asthma. The model developed by Szentpetery *et al.* did not report any of the standard performance metrics in its developmental study, therefore it was not possible to evaluate the generalisability of the model itself. However, the model was still considered in order to promote direct comparison with the other models within this single population.

For the five selected models, data on all predictors were available in the IOWBC, except for race and wheeze apart from cold. As described in Chapter 2.1.1.1, data on frequent wheeze was used as a surrogate variable for wheeze apart from cold. Similarly, a Caucasian ethnicity was assumed for all individuals in order to prevent a substantial loss in sample size. Genotype data in the IOWBC confirmed that the prevalence of non-Caucasian individuals was <2%, therefore this assumption is unlikely to have biased the analysis.

Generalisability could be evaluated for three models – model performance reported in the developmental studies (original studies) and the current validation in the IOWBC varied depending on the performance metric evaluated (Table 4.1). Sensitivity was consistently lower upon implementation of the models in the IOWBC, yet specificity remained comparable, if not higher, upon validation. The ucAPI validated poorly - validation performance was either poorer (5 performance measures) or comparable (2 performance measures) compared to its developmental performance. In contrast, the uPIAMA and PARS models validated well, with comparable AUC with their original studies (0.75 vs. 0.75 and 0.77 vs. 0.80, respectively), and improvements in validation performance across four performance measures each (Figure 4.1, Table 4.1).

Comparing between the five models implemented in the IOWBC, the IoW model demonstrated similar discriminative ability (AUC=0.73) to the uPIAMA and PARS models. Predictors of parental asthma and SPTs were common between these models (detailed in Chapter 3.3.1.2). In contrast, the ucAPI and Szentpetery *et al*. risk score only offered modest discrimination (AUC=0.59 and 0.63, respectively).

The IoW model outperformed the other four models in terms of accuracy, specificity, PPV and positive likelihood ratio, yet it had the lowest model sensitivity (15%). The PARS model demonstrated superior predictive performance for the remaining performance measures, including AUC (Table 4.1).

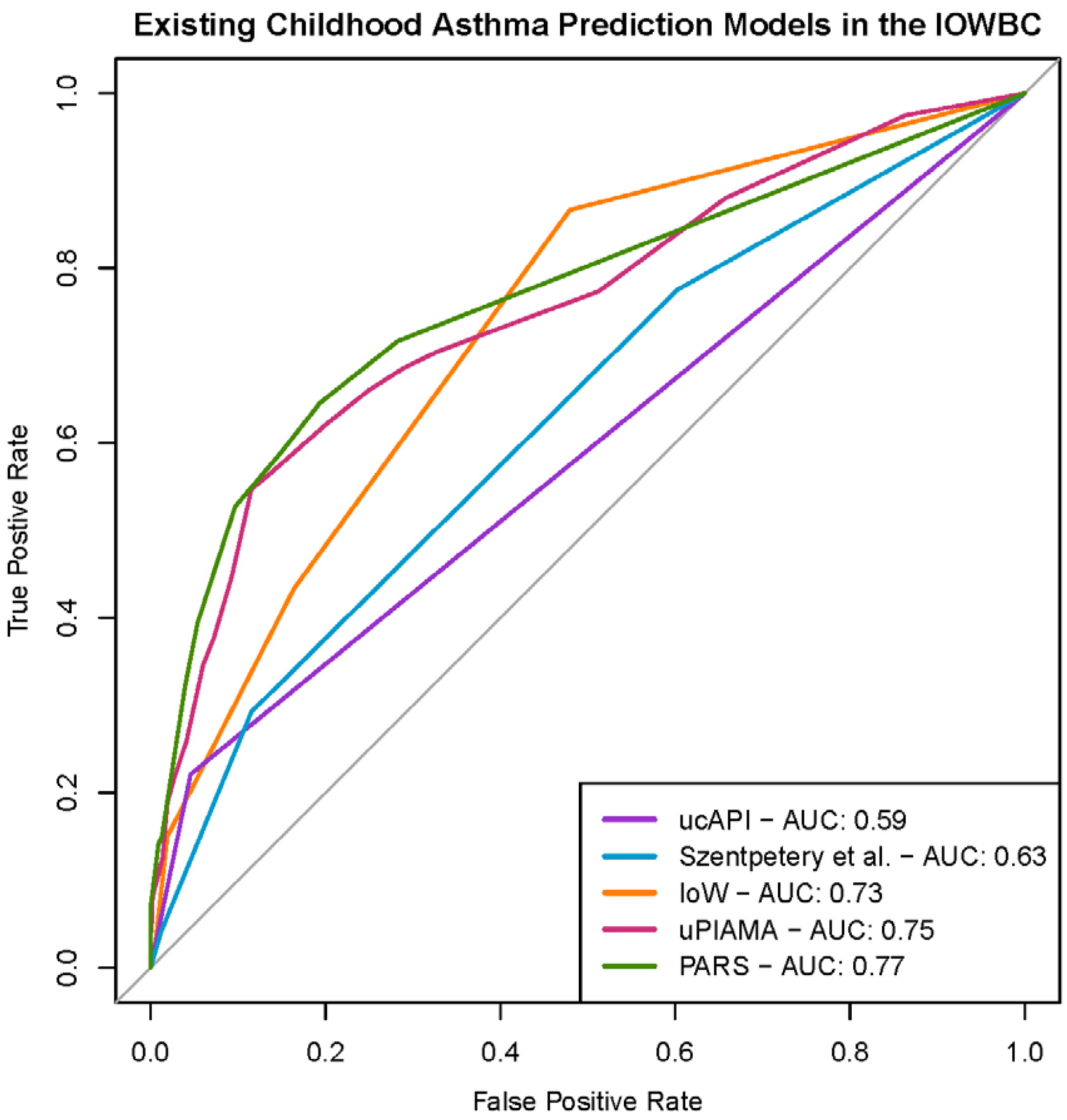


Figure 4.1 ROC curves comparing the performance of existing childhood asthma prediction models in the IOWBC

Table 4.1 Summary of the performance of existing childhood asthma prediction models validated in the single population of the IOWBC

	ucAPI		**Updated PIAMA**		**Szentpetery *et al.***		**PARS**		**IoW**	
	Original [220]	**IOWBC (n=740)**	**Original** [225]	**IOWBC (n=1045)**	**Original** [241]	**IOWBC (n=998)**	**Original** [234]	**IOWBC (n=913)**	**Original** [235]	**IOWBC (n=804)**
AUC	-	0.59	0.75	0.75	-	0.63	0.80	**0.77**	-	0.73
Accuracy	-	0.85	-	0.82	-	0.85	-	0.85	-	**0.86**
Balanced Accuracy	-	0.59	-	0.67	-	0.51	-	**0.72**	-	0.57
Sensitivity	0.44	0.22	0.64	0.42	-	0.04	0.68	**0.53**	0.53	0.15
Specificity	0.94	0.95	0.74	0.92	-	0.99	0.77	0.90	0.85	**0.98**
PPV	0.60	0.44	0.12	0.57	-	0.38	0.37	0.47	0.68	**0.58**
NPV	0.89	0.88	0.97	0.86	-	0.86	0.93	**0.92**	0.74	0.87
LR+	7.50	4.85	2.44	5.13	-	3.47	3.02	5.46	3.41 [a]	**7.87**
LR-	0.60	0.82	0.49	0.63	-	0.97	0.41	**0.52**	0.56 [a]	0.57
F_1 score	0.51 [a]	0.29	0.20 [a]	0.49	-	0.07	0.48 [a]	**0.50**	0.59 [a]	0.24

The performance of existing childhood asthma prediction models were compared between what was reported in their developmental populations and upon validation in the IOWBC. Values in bold identify the model that offered the highest performance for each performance measure.

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

[a] Where unreported, likelihood ratios were calculated based on reported sensitivity and specificity as: LR+ = sensitivity/ 1- specificity, LR- = 1- sensitivity/ specificity; F_1 score was calculated based on reported sensitivity and precision (PPV) as: F_1 score=2*((precision*recall)/ (precision + recall)).

- Performance not reported

4.4 Discussion

Only five of the existing 21 childhood asthma prediction models were able to be implemented in the IOWBC, with four being external validations in an independent population from which the model was developed. The poor applicability of the remaining 15 models is a direct result of the unavailability of predictor information.

There is an argument for the unavailability of predictor data being a limitation of the chosen validation population rather than being a sign of poor model generalisability. It is important to note that the subjects of the IOWBC were recruited in 1989, and the IOWBC is therefore the earliest established European birth cohort focusing on research into allergic diseases[170]. The International Study of Asthma and Allergies in Childhood (ISAAC) only outlined standardised methodologies and questionnaires, aimed at promoting consistent asthma and allergy research worldwide, in 1995[171]. Standardised ISAAC questionnaires, were only available for implementation at the time of the 10-year follow-up for the IOWBC F_1 cohort, but were available from the initiation of the IoW 3rd Generation (F_2) cohort. For example, wheeze apart from cold is an example of data recommended for collection by the ISAAC questionnaire – this variable was unavailable at the early IOWBC follow-ups but was available in the F_2 cohort, prompting a chi-squared test of association to be conducted between the variable itself and its surrogate variable, frequent wheeze (detailed in Chapter 2.1.1.1). Younger cohorts widely utilise these guidelines and often collect similar data. Therefore, despite the early IOWBC data collection questionnaires being similar to the later recommended ISAAC questionnaires, inconsistencies in data collection may account for some of the data unavailability evident in this analysis. However, missing predictors were often those that required additional medical tests, with the absence of data on blood eosinophilia hindering the replication of most models. Only one model was not replicable due to the absence of phenotypic information alone[237].

Despite being indicative of future severe exacerbations and poor asthma control, blood eosinophilia is not regularly measured in children, particularly within primary care settings[254]. Hence, the poor applicability of existing models in the IOWBC is more likely a result of the inclusion of predictor data not routinely collected in clinical or research settings. Whilst additional clinical testing or the inclusion of biomarkers in predictive models may improve prediction accuracy, at present, a number of models developed using only easily available clinical predictors have shown competitive predictive performance (detailed in Chapter 3). As a result, models that require information collected through highly specialised, or non-routine testing have repeatedly been criticised for their poor potential for widespread clinical application, regardless of the predictive benefit they may possess. This analysis reinforces the idea that biomarkers should demonstrate significant predictive benefit in order to be considered for inclusion in a model.

Four of the five models implemented in the IOWBC can be considered to be external validation analyses. However, it was not possible to assess the generalisability of the model developed by Szentpetery *et al*. as none of the standard performance measures were reported in their developmental study for comparison. This highlights the importance of reporting commonly evaluated metrics during the development of a model, particularly since the assessment of generalisability is crucial for determining the future clinical viability of a model. The ucAPI demonstrated poorer performance upon validation compared to its developmental study - a common conclusion of external validation studies, often due to variations in predictor and outcome definitions and underlying population characteristics between development and validation studies. Conversely, both the uPIAMA and PARS models demonstrated good overall generalisability in the IOWBC despite poorer performance reported for some performance measures.

Existing childhood asthma prediction models were replicated in the IOWBC in order to directly compare model performance, having standardised any study or population-effects that may influence the performance of a model. The replication analysis identified that in the IOWBC, the uPIAMA, PARS and IoW models demonstrate similar predictive performance. Interestingly, all three models included predictors of parental asthma and two included SPTs (IoW and PARS). This reinforces the predictive potential of family history and atopy as risk factors for the development of childhood asthma (discussed in Chapter 1.1.6). However, both of these predictors were included in the ucAPI, and SPT was also a predictor in the model developed by Szentpetery *et al*. As these models demonstrated comparatively poorer performance, the inclusion of these predictors alone cannot foreshadow the predictive potential of a model. Notably, all models performed with poor/moderate sensitivity and PPV. For each performance measure evaluated, the best performance was demonstrated by either the IoW or PARS models. Given that the IOWBC was the developmental population of the IoW model, it is unsurprising that the model demonstrated superior predictive performance compared to the other models which were developed in different populations. Hence, the apparent superior performance of the IoW model demonstrated in this analysis should be treated with caution. In contrast, the PARS model, which was developed in Cincinnati, USA, demonstrated the best predictive performance based on a number of performance measures compared to the other models. This comparative analysis suggests that the PARS model has good generalisability and is superior to the other four models when implemented in the IOWBC.

4.5 Conclusion

This analysis provides an initial single-population comparison of existing childhood asthma prediction models. However, the analysis was limited by the inability to implement 15 of the 21 models due to predictor unavailability, and the fact that the chosen population was not independent from the developmental population of the IoW model. Potential to conduct this analysis in MAAS (for which data was available in this thesis) was also hindered by the absence of key predictor data, such as blood eosinophilia, and one of the 21 prediction models (MAAS-APT) also being developed using that cohort. The process of validating all models in a single population needs to be conducted in a population that has data available for a greater proportion of the 38 predictors used amongst the 21 models, and which has not been used to develop any of the existing models, in order to perform a more comprehensive and objective comparison.

Chapter 5 Development of Childhood Asthma Prediction Models using Machine Learning

5.1 Introduction

Due to the highly heterogeneous nature of childhood asthma (previously discussed in Chapter 1), both under-diagnosis and over-treatment of asthma are common, particularly in early life[112,113]. The ability to predict the development of school-age asthma can help to identify high-risk preschool children and distinguish them from children whose symptoms are likely to be transient[38]. Furthermore, early prediction of asthma susceptibility will be critical for the successful implementation of potential primary prevention strategies to reduce the risk of developing asthma. The systematic review conducted in Chapter 3 identified twenty-one regression-based models for predicting childhood asthma[109]. However, none of these models have been implemented into standard clinical practice, possibly due to relatively weak predictive power, poor (or unknown) generalisability and the need for specialised clinical testing for gene expression and VOCs in exhaled breath condensates. The review further proposed that regression-based methods for predicting childhood asthma may have been exhausted, with the identified models offering similar predictive power to each other and being unable to be significantly improved upon.

Machine learning approaches have increasingly been applied to a wide range of healthcare problems due to their ability to integrate large quantities of heterogeneous data, handle complex interactions between variables and identify patterns within data[118]. Particularly for disease prediction, where interactions between biological variables are complex, machine learning approaches have the potential to identify novel predictors which may have been previously overlooked by regression-based approaches[117,118,123]. Furthermore, application of methods to reduce model overfitting may address the poor generalisability of existing prediction models in independent populations. Machine learning approaches have shown promise in predicting a variety of clinical asthma outcomes, phenotypes and decisions[105,155,158,159,255], including the diagnostic or prognostic prediction of school-age asthma development[161,163-166,251,256-258]. While these studies tend to offer improved predictive performance, none of these studies support their findings with external validations of their models or explain how their “black-box” models (where relevant) arrive at their predictions. Without these two components, machine learning models will fail to obtain the trust of physicians and continue to be limited in their clinical utility, regardless of the superior prediction accuracy they may offer[167,250] (discussed in Chapter 1.3).

5.1.1 Objectives

In this chapter, machine learning approaches are utilised in an attempt to improve upon the performance of traditional regression methods and develop explainable and independently validated prediction models for childhood asthma. Two prognostic prediction models, the Childhood Asthma Prediction in Early-life (CAPE) and Childhood Asthma Prediction at Preschool-age (CAPP) models, were developed to predict school-age asthma at 10 years, within the general population of the IOWBC, using information available within the first two years and first four years of life, respectively.

In line with the aims of this thesis (detailed in Chapter 1.4), the optimal subset of clinical features predictive of school-age asthma were identified for each model (Aim 3). Using these selected features, a number of machine learning models were constructed and compared in order to identify the optimal machine learning algorithm for predicting childhood asthma (Aim 4). Finally, global and local interpretations of the final machine learning models were made in an attempt to explain how predictions were deduced (Aim 5).

5.2 Methods

5.2.1 Developmental study population

The IOWBC (n=1456) was used as the developmental population for constructing the CAPE and CAPP models, (described in Chapter 2.1.1). For both models, 1368 participants with a defined prediction outcome of school-age asthma were included in the analyses (detailed in Chapter 2.2).

As described in Chapter 2.1.1.1, data for 54 candidate predictors extracted from the IOWBC were considered during the development of the clinical prediction models (Table A1). Only candidate predictor data available from the birth and early-life time-points were considered for the development of the CAPE model whilst predictors across all three time-points (birth, early life and 4-year follow-up) were considered for the development of the CAPP model.

Pre-processing of candidate predictor data was performed to prepare the data for model construction. Potentially extreme outliers (±4SD) present among the continuous variables were removed. Further transformation of continuous data was not required as histograms for each continuous variable demonstrated a Gaussian data distribution (Figure A2). One-hot-encoding was used to encode categorical variables without an ordinal interpretation (nominal variables) into separate binary variables.

All stages of model development were performed independently for the CAPE and CAPP models (Figure 5.1).

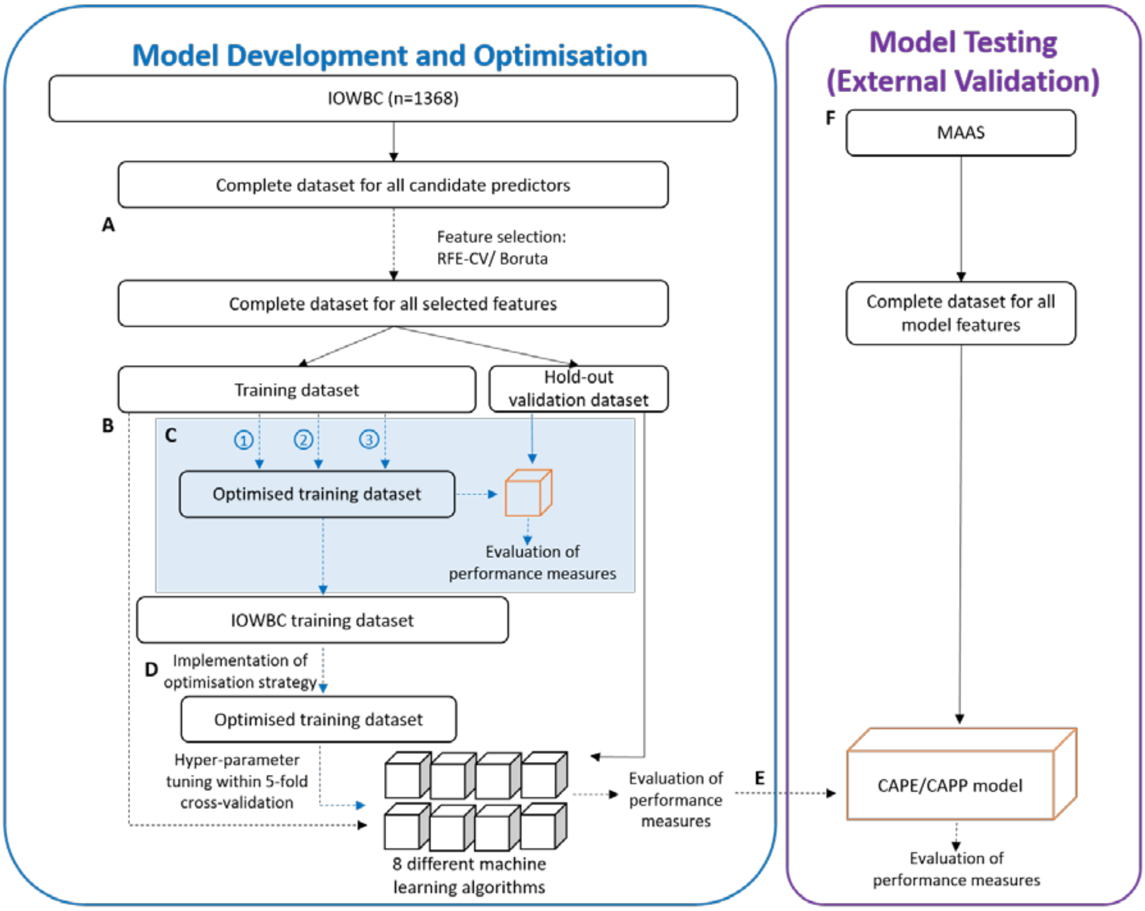

Model Development and Optimisation
IOWBC (n=1368)
Complete dataset for all candidate predictors
A
Feature selection:
RFE-CV/ Boruta
Complete dataset for all selected features
Training dataset
Hold-out validation dataset
B
C
①
②
③
Optimised training dataset
Evaluation of performance measures
IOWBC training dataset
D
Implementation of optimisation strategy
Optimised training dataset
Hyper-parameter tuning within 5-fold cross-validation
8 different machine learning algorithms
Evaluation of performance measures
E
Model Testing (External Validation)
F
MAAS
Complete dataset for all model features
CAPE/CAPP model
Evaluation of performance measures

Figure 5.1 Workflow for the development and validation of the CAPE and CAPP models

Models were developed using data from the Isle of Wight Birth Cohort (IOWBC), n=1368 (14.7% asthmatic). (A) Feature selection was performed using only individuals with complete data for all candidate predictors. (B) Models, using eight machine learning classifiers, were first developed using only individuals with complete data for the subset of features identified from feature selection. (C) Three training process optimisation strategies were assessed on the best performing model identified at (B): 1) imputation, 2) oversampling and 3) random undersampling. (D) All combinations of optimisation strategies were applied to the dataset of all individuals in the IOWBC not allocated to the validation dataset (IOWBC training dataset) in a step-wise approach – n=1113 (15.0% asthmatic) for the early life training dataset and n=1185 (14.9% asthmatic) for the preschool training dataset. Models were redeveloped for all algorithms using each optimised training datasets. (E) The best models for use in early life (CAPE tool) and at preschool age (CAPP tool) were selected based on performance in the holdout validation set. (F) Selected models were externally validated in the independent population of the Manchester Asthma and Allergy Study (MAAS).

5.2.2 Feature selection

For each model, feature selection was performed on the complete dataset of all available candidate predictors (without any missing values). Two wrapper feature selection methods were compared - recursive feature elimination within a five-fold cross-validation (RFECV) (described in Chapter 2.3.1.1), and Boruta (described in Chapter 2.3.1.2). Both methods utilised a variation of the random forest algorithm (described in Chapter 2.3.1). Continuous variables were standardised to zero mean and unit variance prior to feature selection.

The best feature selection method was chosen based on a combination of factors. First, the subsets of selected features were evaluated for their biological plausibility and complexity for model construction. Next, the ability of each method to handle the potential presence of multicollinearity was assessed by evaluating the presence of highly correlated features within the selected feature subsets (based on their Pearson’s correlation coefficient). Feature importance was assessed using the feature importance attribute of the random forest algorithm. Further insight into the direction and magnitude of risk associated with each predictor was then extracted using SHAP (detailed in Chapter 2.3.7).

5.2.3 Complete-case machine learning models development

As described in Chapter 2.3.2, eight supervised machine-learning classifiers were compared to identify the best algorithm for this classification problem: three SVMs with different kernel functions, decision tree, random forest, naive Bayes classifier, MLP, and KNN. The models were initially developed using data on individuals with complete data for the subset of features identified through the feature selection process (Figure 5.1, Stage B).

Independently, for the CAPE and CAPP models, the complete dataset was split (ratio of 2:1, preserving class proportions) into a training and holdout validation dataset for model development and evaluation, respectively. The continuous variables in the training data were standardised to a mean of zero and unit variance, and the same standardisation properties applied to the holdout validation set. Within the training process, the hyperparameters for each model were tuned using a grid search, within a 5-fold cross validation (detailed in Chapter 2.3.3). For each model, the set of hyperparameters with the highest average balanced accuracy score across the folds was selected. The optimal tuning parameters were used to train each model on the entire training set. The trained models were then used to make predictions on the holdout validation set. Performance measures for each model (detailed in Chapter 2.3.6), were reported for both the training and validation datasets.

5.2.4 Exploration of model optimisation techniques

Three optimisation techniques of imputation, oversampling and undersampling were implemented to address the issues of missing data and class imbalance that were observed in the early life and preschool training datasets. The best performing algorithm identified from the complete data model development stage was used to assess whether each optimisation technique offered any predictive benefit (Figure 5.1, Stage C). The early life and preschool holdout validation datasets were not modified, thus remained as single unseen complete datasets that could be used to evaluate and compare the predictive performance of each model, trained on either the complete or optimised training datasets.

5.2.4.1 Imputation of missing data

To increase the number of training examples and reduce potential biases introduced through the complete data analyses, imputation was performed. Two imputation methods, missForest (detailed in Chapter 2.3.4.1) and MICE (detailed in Chapter 2.3.4.2), were compared. Missing data was randomly introduced to the complete training dataset (20% of the dataset). The missing values were then imputed using the two imputation methods. After the implementation of each imputation strategy, the continuous variables in the training dataset were standardised to a mean of zero and unit variance, and the same standardisation properties applied to the test dataset. The best performing machine learning algorithm identified from the complete data model development stage was then retrained on each imputed training dataset. The best imputation method was determined based on the model that offered the most similar performance to the model trained using the complete training dataset.

5.2.4.2 Handling the class imbalance

To reduce the class imbalance observed in the early life and preschool training datasets, which was skewed towards controls (1:7 and 1:6 case: control ratio, respectively), oversampling using ADASYN was implemented (detailed in Chapter 2.3.5.1). To identify the optimal degree of oversampling required, seven levels of oversampling were evaluated, increasing the number of the minority class (asthma cases) by 25%, 50%, 100%, 150%, 200%, 250% and 300%.

The impact of oversampling in addition to random undersampling of the majority class (non-asthmatic controls) was also assessed (detailed in Chapter 2.3.5.2). The aim of combining both oversampling and undersampling was to obtain a balanced training dataset, with a 1:1 ratio of asthmatic and non-asthmatic examples.

5.2.5 Implementation of optimisation strategies for final model development

Each optimisation technique was considered for widespread implementation if it demonstrated improvements in the predictive performance of the best performing model chosen from the complete model development stage. Upon signs of improvement, the optimisation techniques were applied to the training datasets across all possible combinations: i) complete training data with oversampling; ii) complete training data with undersampling; iii) complete training data with oversampling and undersampling; iv) imputed training data; v) imputed training data with oversampling; vi) imputed training data with undersampling; and vii) imputed training data with oversampling and undersampling.

Specifically, if imputation was deemed beneficial, the best imputation method was first used to impute missing data for all 1368 individuals not allocated to the holdout validation dataset for the early life and preschool models, independently. The continuous variables in the imputed training dataset were then standardised to a mean of zero and unit variance, and the same standardisation properties were applied to the holdout validation dataset. Next, if oversampling demonstrated a predictive improvement, ADASYN was applied to the complete or imputed training dataset across all seven levels of oversampling previously evaluated. Finally, if shown to offer a predictive improvement, random undersampling was applied to each training dataset to obtain a 1:1 class balance. The eight machine learning algorithms were then redeveloped and tuned on each of the optimised training datasets (as previously described) in order to identify the best CAPE and CAPP models (Figure 5.1, Stage D).

5.2.6 Evaluation of model performance

Performance measures of discrimination, sensitivity, specificity, positive and negative predictive values (PPV and NPV, respectively), positive and negative likelihood ratios (LR+ and LR-, respectively), balanced accuracy and F_1-score were reported for each machine learning algorithm in both the training and validation datasets (defined in Chapter 2.3.6).

The best CAPE and CAPP models were selected based on: i) their discriminative performance in the validation set; and ii) a judgement on any potential signs of overfitting (large differences in model performance reported for the training and test sets). Although the ability to correctly identify both future asthmatics and non-asthmatics is important, the expert opinion of a consultant respiratory physician and research professor suggested that the ability for a model to detect true future asthmatics was preferred over identifying true non-asthmatics. Therefore, where the selection of the best model based on the above criteria was not clear, performance measures indicating the ability to rule in asthma cases (e.g. sensitivity, PPV and F_1-score) were also considered (Figure 5.1, Stage E).

Upon selection of the final CAPE and CAPP models, performance in the validation and test (MAAS) sets were reported at the optimal threshold that maximised the Youden's Index, with 2000 bootstrap samples used to calculate 95% confidence intervals for the performance measures. The Brier score was also calculated for these selected models.

5.2.7 External validation

The generalisability of the CAPE and CAPP models was assessed through an external validation using the MAAS cohort (detailed in Chapter 2.1.2) (Figure 5.1, Stage F). In this independent test set, data for model predictors and the asthma outcome (evaluated at age 8 and 11 years) were closely matched to maximise the similarity of the definitions used in the IOWBC (Table A3). Data cleaning also corresponded to processes conducted in the developmental cohort (detailed in Chapter 5.2.1). Only individuals with complete data for the predictors and outcome were used in the external validation analyses.

Model generalisability was assessed in MAAS among three risk groups – unselected, moderate and high risk based on a parental history of allergic disease (asthma, eczema or allergic rhinitis), with zero, one or two parents affected, respectively.

5.2.8 Sensitivity analyses

Sensitivity analyses were conducted to comprehensively evaluate the CAPE and CAPP models, including evaluations of: i) the robustness to predict an alternative definition of school-age asthma; ii) the resolution of the predictions to identify individuals presenting with distinct wheeze phenotypes throughout childhood; and iii) the performance of the machine learning models compared to similar regression-based models.

5.2.8.1 Assessing the robustness to predict an alternative asthma definition

The robustness of the CAPE and CAPP models was evaluated using an alternative definition of school-age asthma that incorporated an objective outcome measure. Using this alternative asthma definition, a child was considered asthmatic if they presented with wheeze in the last 12 months and had bronchial hyper-responsiveness (BHR) (defined in Chapter 2.1.1.1). Due to the requirement for both the presence of wheeze and positive BHR to confirm asthma at age 10, individuals who were reported to have no wheeze at age 10, irrespective of whether BHR was assessed, could be assigned as non-asthmatic. As such, although BHR was only assessed in 784 individuals in the IOWBC, an alternative asthma status could be obtained for 1312 individuals.

5.2.8.2 Ability to predict distinct wheeze phenotypes

The resolution of the asthma predictions to distinguish between individuals presenting with distinct wheeze phenotypes throughout childhood and adolescence was assessed. The identification and assignment of individuals in the IOWBC and MAAS into one of five distinct wheeze phenotypes through a latent class analysis has previously been described[34]. Briefly, using wheeze data available across five time-points, a latent class analysis of 7,719 individuals from five UK birth cohorts (including the IOWBC and MAAS) identified five distinct phenotypes of wheeze: never/infrequent wheeze, early onset preschool remitting, early onset mid-childhood remitting, persistent, and late-onset wheeze (full details on the analysis can be found in reference [34]). For each individual, the latent class analysis provided probabilities of belonging to each wheeze phenotype. In this analysis, each individual was categorised to their most probable wheeze phenotype. The ability for the models to predict these distinct wheeze phenotypes was then assessed based on the proportion of individuals offered a positive asthma prediction in each phenotype group.

5.2.8.3 Comparison with regression-based methods

To evaluate the hypothesis that machine learning methods may offer more accurate predictions of childhood asthma than regression-based methods, the CAPE and CAPP machine learning models were directly compared with equivalent logistic regression models developed using the same predictors. Where data was available in the IOWBC, the performance of the machine learning models were also compared against their existing regression-based benchmark models.

5.2.8.3.1 Derivation of equivalent logistic regression models for the CAPE and CAPP models

To directly compare if the use of more complex machine learning algorithms can offer more accurate prediction of childhood asthma than regression-based methods, CAPE and CAPP equivalent logistic regression models were constructed. The logistic regression equivalent models were developed using the same predictors and training datasets used to construct the CAPE and CAPP machine learning models. To construct the logistic regression models, the scikit-learn logistic regression algorithm was used ('lbfgs' solver)[127]. No further regularisation of the predictor coefficients was applied during the construction of the logistic regression models in order to support direct comparison with previously published work.

5.2.8.3.2 Comparison of the CAPE and CAPP models with current benchmark models

The developed machine learning models were further compared against currently published models. The API, the most widely known asthma prediction tool, was unable to be replicated due to the absence of data on blood eosinophil counts in the IOWBC. Of the remaining validated models, the PAPS (Persistent Asthma Predictive Score)[232] and PARS (Paediatric Asthma Risk Score)[234] were considered the best performing models comparable with the CAPE and CAPP models, offering predictions in early life and at preschool age, respectively. However, PAPS was also unable to be replicated as specific IgE tests were not performed in the IOWBC. PARS was able to be replicated in both the IOWBC and MAAS.

Replication of the PARS model in the IOWBC is detailed in Chapter 4, and the same method was applied in MAAS (predicting asthma in the IOWBC at age 10: n=913, in MAAS at age 8 years: n=552, in MAAS at age 11 years: n=487). First, the performance of the PARS model was compared against CAPP based on AUC. Next, among the subset of individuals with predictions from both the CAPP and PARS models, differences in predictions made by the two models were compared. In the IOWBC, only individuals in the test set (i.e. not used to train the model) were assessed. Reclassification tables were used to evaluate the differences in predictions on an individual level, for asthmatics and non-asthmatic individuals separately[259]. The tables present the differences in prediction classification using the CAPP model compared to the PARS model. The net proportion of individuals reclassified by the CAPP model to a more appropriate prediction group was summarized by the net reclassification indices for true future asthmatics and non-asthmatics separately (NRI_{event} and $NRI_{non\text{-}event}$, respectively, Equation 5.1)[259]:

$$NRI_{event} = P(up|event) - P(down|event)$$

$$NRI_{non\ event} = P(down|non\ event) - P(up|non\ event)$$

Equation 5.1 Calculations to deduce the net reclassification indices for events and non-events

In the context of this thesis, an event=asthmatic, non-event=non-asthmatic, up=reclassified as asthmatic; down=reclassified as non-asthmatic.

5.2.9 Explaining the “black-box” models

SHapley Additive exPlanations (SHAP) (detailed in Chapter 2.3.7) were used to evaluate feature importance and provide global explanations for how predictions were made by the CAPE and CAPP models. Based on the global explanations of each model, the CAPE and CAPP models were redeveloped using the subset of features shown to offer the greatest contribution to the model predictions. Examples of using SHAP to explain individual predictions were also provided.

5.3 Results

In the IOWBC, 1368 individuals had an asthma outcome at age 10 and were included in the study (prevalence of asthma at age 10 was 14.69%). Based on data available in the IOWBC, 54 candidate predictors corresponding to known risk factors of childhood asthma were extracted across three time-points – birth, early life (1 and 2-year follow-ups), and preschool age (four-year follow-up). Descriptive statistics and hypothesis testing identified that asthmatic children were significantly more likely to be male, have a lower birthweight, be atopic and experience asthma-like symptoms both in early life and at preschool age compared to non-asthmatic children in the IOWBC at age 10 (Table A4).

5.3.1 Comparability between the IOWBC and datasets used for model development

Feature selection was performed on individuals with complete data for all candidate features for both the early life (n=490 with 39 predictors) and preschool (n=373 with 54 predictors) models. Asthma prevalence among those with complete data for each group (14.29% and 14.75%, respectively) was similar to the 1368 individuals analysed in the IOWBC (14.69%).

Both the early life and preschool complete datasets were largely comparable with the total IOWBC analysed, based on the hypothesis tests used to compare differences between asthmatic and non-asthmatic children in each dataset (Table A4). However, the statistically significant differences in gender (X^2=5.41, p-value=0.02), birthweight (X^2=2.49, p-value=0.01) and early life polysensitisation (X^2=24.23, p-value<0.01), which were observed between asthmatic and non-asthmatic children in the total IOWBC analysed, were not observed in either the early life or preschool complete datasets. Similarly, in contrast to the total IOWBC analysed, the complete preschool dataset showed no statistically significant difference between asthmatic and non-asthmatic children in terms of eczema or monosensitisation evaluated at age 4. However, as the same trends were reflected in the complete datasets (except birthweight in the preschool complete-case dataset), it is likely that the lack of significance observed was due to low statistical power. Additionally, in contrast to the total IOWBC analysed, there was a statistically significant difference for a history of paternal eczema in the early life complete dataset (X^2=5.08, p-value=0.02). Notably, in both the early life and preschool complete datasets, the proportion of asthmatic children with a reported history of maternal asthma was substantially lower than that observed in the total IOWBC (7.14%, 5.45% and 14.43%, respectively); however, no significant difference was observed between the asthmatic and non-asthmatic individuals in the original IOWBC, the early life or preschool complete datasets (Table A4).

The correlation matrix constructed to assess the candidate features for the presence of multicollinearity showed that the majority of features were uncorrelated (Figure A3). Two main clusters of highly correlated features were evident. The features within these clusters were related to the presence of asthma-like symptoms at the early life and preschool time-points, suggesting that individuals demonstrating asthma-like symptoms were likely to have presented with multiple symptoms. As expected, as frequent wheeze was used as a surrogate for wheeze without cold, these predictors were perfectly correlated (*r*=1, p<0.01). Some of the features that were repeated across multiple time-points were observed to be highly correlated, either positively or negatively. For example, individuals whose parents were non-smokers in early life were also reported as non-smokers when the child was of preschool-age (*r*= 0.93, p<0.01).

5.3.2 Feature selection

Feature selection using the RFECV method, identified optimal subsets of 8 and 12 features, attaining an average cross-validation balanced accuracy score of 65% and 75%, for the early life and preschool models, respectively (Figure 5.2, Table 5.1). Six predictors were common between the early life and preschool model feature subsets. Predictors of wheeze and cough were both represented in the feature subsets, but at the later 4-year follow-up time point for the preschool model. Nocturnal symptoms, atopy and polysensitisation were additional features selected for the preschool model.

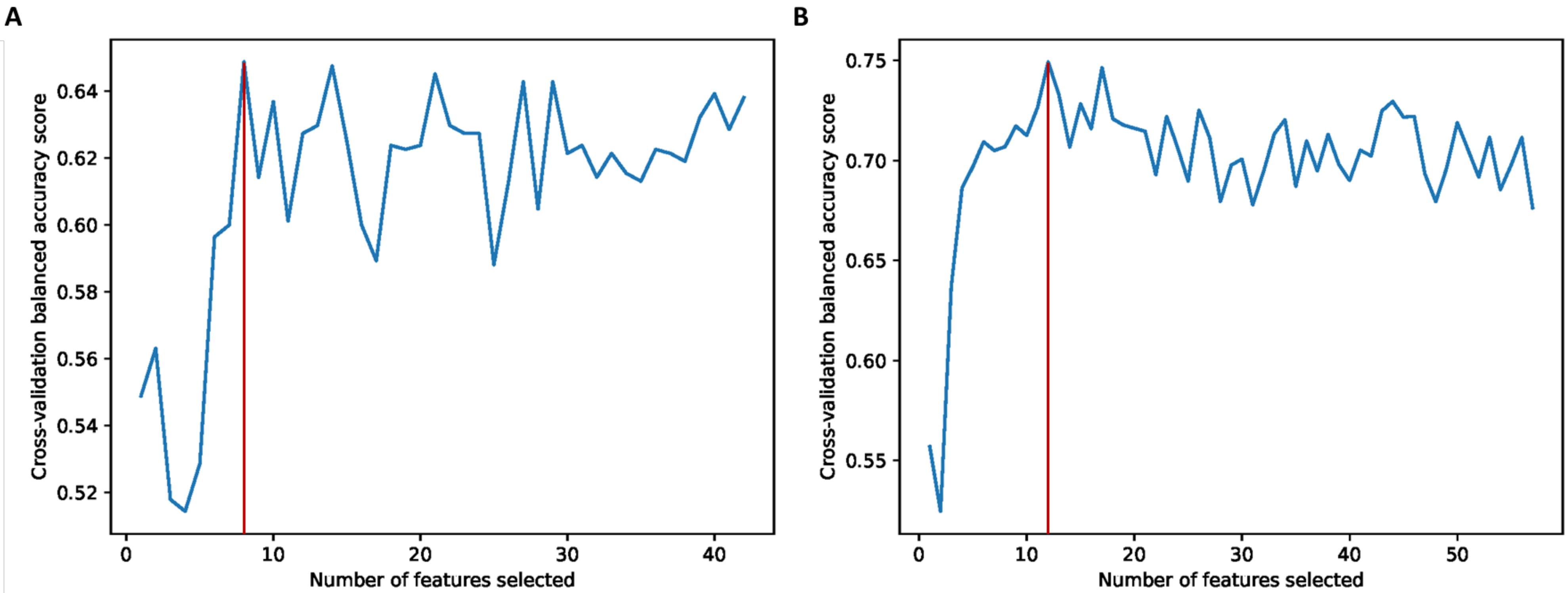


Figure 5.2 Feature selection using Recursive Feature Elimination for the early life and preschool models

The optimal set of predictors for inclusion in the early life (A) and preschool (B) models was identified as the subsets of features that reached the highest cross-validation balanced accuracy (red line).

Table 5.1 Predictors selected for inclusion in the early life and preschool models by RFE

Model	Average balanced accuracy (%)	Features (n)
Early life	64.49	(8) - Maternal age, birthweight, total breastfeeding duration, age of solid food introduction, BMI at 1 year, early life wheeze, early life cough and maternal socioeconomic status
Preschool	74.93	(12) - Maternal age, birthweight, total breastfeeding duration, age of solid food introduction, BMI at 1 year, BMI at 4 years, preschool wheeze, preschool cough, preschool nocturnal symptoms, preschool atopy status, preschool polysensitisation and maternal socioeconomic status

The Boruta feature selection method identified preschool cough as the only predictive feature for the preschool model. For the early life model, no predictive features were selected suggesting that none of the candidate features were better than random variables at predicting school-age asthma.

Based on the comparison between the two feature selection methods, RFE was chosen as the best feature selection method to support the development of both the early life and preschool models. This was primarily due to the lack of features selected as predictive of asthma at age 10 by the Boruta method. The RFE method was also deemed to be a robust feature selection method as the selected feature subsets appeared able to account for collinearity between candidate features. For example, the early life and preschool subsets both identified predictors which were represented across multiple time-points and sometimes presented to be highly correlated with each other; however, when selected, the predictor was only selected at one of the time-points. Only BMI was repeated as a predictor in the preschool model, evaluated at both 1 and 4 years.

Using the in-built feature importance measure of the random forest algorithm, all of the candidate features considered for the early life (Table A5) and preschool models (Table A6) were ranked by their predictive importance. Rather than just identifying the top ranking features (as selected for the early life model), the RFE method appeared to identify a selective subset of important features for the preschool model, which together demonstrated the optimal predictive performance.

SHAP summary plots were generated in an attempt to uncover the contribution of each predictor selected by RFE for inclusion in the early life and preschool models (Figure 5.3). The random forest algorithm and the complete dataset used during RFE were used to generate the plots. Unlike the feature importance attribute of the random forest algorithm, SHAP summary plots provided insight into the importance, direction and magnitude of risk associated with each predictor included in the models. Based on SHAP, the presence of cough was considered the most important predictor in both the early life and preschool models, offering the greatest impact on the final classification. This was followed by the presence of other asthma related symptoms and markers of atopy (the latter for the preschool model only). There is a clear indication that the presence of these predictors contribute towards an individual being predicted as asthmatic. Whilst the direction of risk for the other predictors appear less discrete, they still offer some insight. For example, in both models, a shorter duration of total breastfeeding appears to offer a higher contribution to a prediction of asthma than a longer duration of breastfeeding.

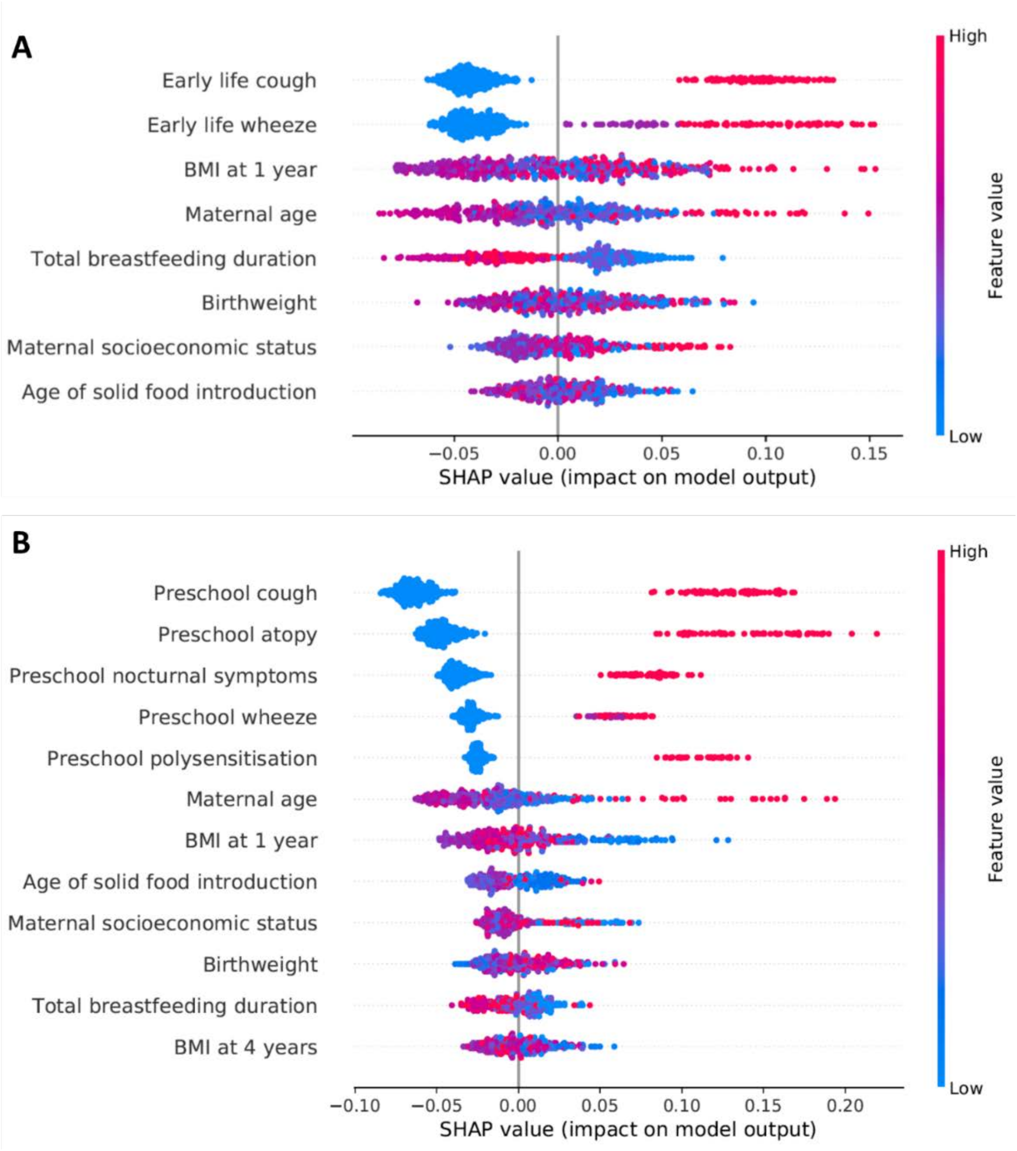


Figure 5.3 SHAP feature importance values for the early life and preschool models

SHAP summary plots of the features selected during RFE for the early life (A) and preschool models (B). Predictors are listed in descending order of their SHAP value. The higher the SHAP value, the larger its contribution (importance) on model predictions. Each dot in each predictor row corresponds to a separate individual. The placing of the dot along the x-axis represents the contribution of the predictor in the individual's asthma prediction. The colour of the dot refers to the feature value, with higher values coloured red and lower values in blue. For example, early life cough offers the highest contribution to the random forest model, with higher values (presence of early life cough) having a positive contribution towards a prediction of asthma. The absence of early life cough (blue dots) reduces its contribution to the model delivering a prediction of asthma.

5.3.3 Complete machine learning model development

Complete data for the eight predictors selected for the CAPE model was available for 765 individuals. Following the stratified train-test split, 510 (68 asthmatics) and 255 (34 asthmatics) individuals were allocated to the initial training and holdout validation sets, respectively. Similarly, complete data for the 12 predictors selected for inclusion in the CAPP model was available for 548 individuals, of whom 365 (51 asthmatics) and 183 (25 asthmatics) individuals were assigned to the initial training and holdout validation sets, respectively.

All eight machine learning algorithms trained on the complete early life training dataset demonstrated modest discriminative performance in the holdout validation set, with AUC ranging between 0.54 and 0.62 (Figure 5.4A, Table 5.2). Whilst the MLP model demonstrated the best performance in terms of AUC and specificity, the model showed significant signs of overfitting, unable to correctly identify any future asthmatic children in the validation set (0% sensitivity, PPV and F_1-score). The random forest model demonstrated the best sensitivity, at 29%. However, the decision tree and KNN models each demonstrated the best performance across four different performance measures. Whilst the decision tree model reported the best balanced accuracy, NPV, LR- and F_1 score, the KNN early life model demonstrated the best specificity, PPV, LR+ and overall accuracy (Table 5.2).

In contrast, the preschool models developed using the complete training dataset demonstrated superior discriminative ability compared to the early life models, with AUC ranging between 0.61 and 0.78 in the holdout validation set (Figure 5.4B, Table 5.3). The KNN model performed with the best discrimination and specificity. Yet, whilst the linear SVM demonstrated equivalent discrimination and comparable specificity (0.92 vs 0.97), its sensitivity was more than double that of the KNN model (0.40 vs 0.16, respectively). The SVM using the RBF kernel, decision tree and random forest models all demonstrated the best performance in terms of PPV and LR+ (0.50 and 6.32, respectively). Although these three models reported equivalent PPV, a PPV of 0.5 indicates that, of those individuals predicted to have school-age asthma by the model, the probability that the individual will actually have asthma is only 50%. In comparison to the other models, the naïve Bayes model correctly predicted the most future asthmatic cases and reported the best performance across five different performance measures (balanced accuracy, sensitivity, NPV, LR- and F_1-score) (Table 5.3).

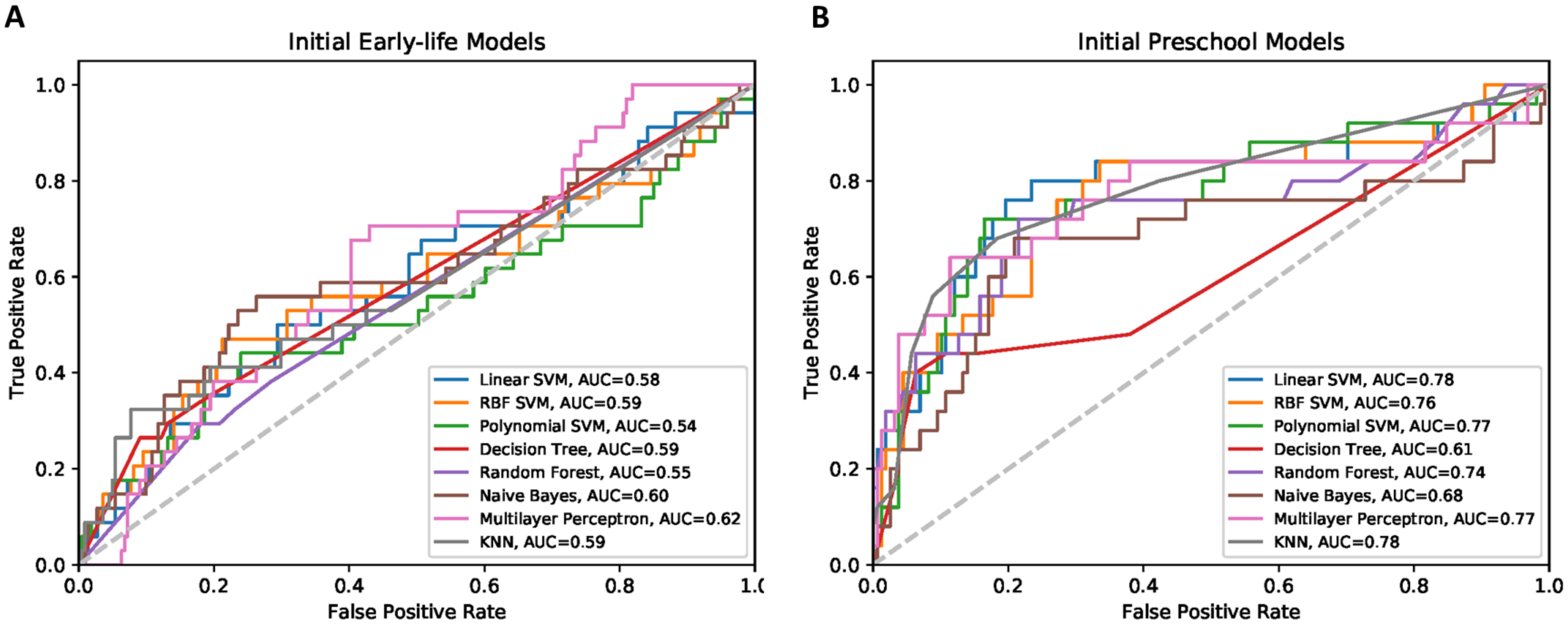


Figure 5.4 ROC curves comparing the performance of the early life and preschool models developed using complete training datasets

Table 5.2 Performance of the eight machine learning algorithms developed using the complete early life training dataset

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
SVM (Linear)	0.87	0.50	0.64	0.00	1.00	-	0.87	-	1.00	-	442, 0, 68, 0
	0.87	0.50	0.58	0.00	1.00	-	0.87	-	1.00	-	221, 0, 34, 0
SVM (RBF)	0.90	0.65	0.89	0.31	1.00	0.91	0.90	68.25	0.69	0.46	440, 2, 47, 21
	0.85	0.55	0.59	0.15	0.96	0.36	0.88	3.61	0.89	0.21	212, 9, 29, 5
SVM (Polynomial)	0.98	0.93	0.97	0.87	1.00	1.00	0.98	-	0.13	0.93	442, 0, 9, 59
	0.78	0.56	0.54	0.26	0.86	0.23	0.88	1.89	0.86	0.24	190, 31, 25, 9
Decision Tree	0.99	0.96	1.00	0.91	1.00	1.00	0.99	-	0.09	0.95	442,0, 6,62
	0.82	**0.59**	0.59	0.26	0.91	0.31	**0.89**	2.93	**0.81**	**0.29**	201,20, 25,9
Random Forest	0.92	0.86	0.88	0.78	0.94	0.65	0.97	12.30	0.24	0.71	414, 28, 15,53
	0.73	0.55	0.56	**0.29**	0.80	0.19	0.88	1.48	0.88	0.23	177, 44, 21,10

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Naïve Bayes	0.84	0.65	0.69	0.40	0.91	0.40	0.91	4.28	0.66	0.40	401, 41, 41, 27
	0.80	0.55	0.60	0.21	0.90	0.23	0.88	1.98	0.89	0.22	198, 23, 27, 7
MLP	0.88	0.56	0.81	0.13	1.00	0.82	0.88	29.25	0.87	0.23	440,2, 59,9
	0.85	0.49	**0.62**	0.00	**0.99**	0.00	0.87	0.00	1.01	0.00	218,3, 34,0
KNN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-	0.00	1.00	442, 0, 0, 68
	0.87	0.54	0.59	0.09	**0.99**	**0.50**	0.88	**6.50**	0.92	0.15	218, 3, 31, 3

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

Shaded rows report performance in the training set, unshaded rows report performance in the holdout validation set, bold= highest model performance.

[a] The final column presents the confusion matrix for the model classifications, where TN=true negatives, FP=false positives, FN=false negatives, TP=true positives.

Table 5.3 Performance of the eight machine learning algorithms developed using the complete preschool training dataset

	Accuracy	**Balanced Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**LR+**	**LR-**	**F_1 Score**	**TN, FP, FN, TP** [a]
SVM (Linear)	0.88	0.67	0.86	0.39	0.96	0.63	0.91	10.26	0.63	0.48	302, 12, 31, 20
	0.85	0.66	**0.78**	0.40	0.92	0.45	0.91	5.27	0.65	0.43	146, 12, 15, 10
SVM (RBF)	0.92	0.71	0.90	0.43	1.00	0.96	0.92	135.45	0.57	0.59	312, 2, 29, 22
	0.86	0.60	0.76	0.24	0.96	**0.50**	0.89	**6.32**	0.79	0.32	152, 6, 19, 6
SVM (Polynomial)	0.98	0.92	0.99	0.84	1.00	1.00	0.98	-	0.16	0.91	314, 0, 8, 43
	0.83	0.70	0.77	0.52	0.88	0.41	0.92	4.32	0.55	0.46	139, 19, 12, 13
Decision Tree	0.94	0.79	0.97	0.59	0.99	0.94	0.94	92.35	0.41	0.72	312, 2, 21, 30
	0.86	0.60	0.61	0.24	0.96	**0.50**	0.89	**6.32**	0.79	0.32	152, 6, 19, 6
Random Forest	0.92	0.73	0.90	0.47	0.99	0.89	0.92	49.25	0.53	0.62	311, 3, 27, 24
	0.86	0.65	0.74	0.36	0.94	**0.50**	0.90	**6.32**	0.68	0.42	149, 9, 16, 9

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F$_1$ Score	TN, FP, FN, TP [a]
Naïve Bayes	0.86	0.78	0.83	0.67	0.89	0.50	094	6.16	0.37	0.57	280, 34, 17, 34
	0.83	**0.76**	0.68	**0.68**	0.85	0.41	**0.94**	4.48	**0.38**	**0.52**	134, 24, 8, 17
MLP	0.88	0.74	0.84	0.55	0.93	0.57	0.93	8.21	0.48	0.56	293, 21, 23, 28
	0.84	0.71	0.77	0.52	0.89	0.43	0.92	4.83	0.54	0.47	141, 17, 12, 13
KNN	0.90	0.65	0.91	0.31	0.99	0.89	0.90	49.25	0.69	0.46	312, 2, 35, 16
	0.86	0.56	**0.78**	0.16	**0.97**	0.44	0.88	5.06	0.87	0.24	153, 5, 21, 4

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

Shaded rows report performance in the training set, unshaded rows report performance in the holdout validation set, bold= highest model performance.

[a] The final column presents the confusion matrix for the model classifications, where TN=true negatives, FP=false positives, FN=false negatives, TP=true positives.

5.3.4 Exploration of model optimisation techniques

Based on the performance of the early life and preschool models developed using complete training datasets, the best performing model was selected to assess whether addressing the extent of missing data and class imbalance present in the training datasets could improve the predictive performance of the developed models. AUC was the main criteria used to select this best performing model. Whilst the preschool linear SVM and KNN models offered equivalent AUC, the preschool linear SVM was chosen due its superior sensitivity.

5.3.4.1 Imputation

The missForest and MICE imputation methods were first compared on the complete preschool training dataset, into which 20% of missing data was randomly introduced. Of the two imputation methods compared, the linear SVM developed using the MICE imputed training dataset performed most similarly to the linear SVM trained on the complete preschool training dataset – classifications on the holdout validation set only differed by one false-positive prediction (Table 5.4). All performance measures generated through the MICE-imputed strategy were either equivalent or marginally inferior to those reported based on the complete data strategy.

Interestingly, compared to the linear SVM developed using the complete training dataset (observed real-world data), the missForest imputed training dataset (containing estimates for 20% of the same training dataset) was better able to correctly predict true asthmatics in the validation set (also observed real-world data). Furthermore, the linear SVM developed using the missForest imputation method demonstrated superior predictive performance in terms of balanced accuracy, AUC, sensitivity, NPV, LR- and F_1-score compared to both of the models developed using the complete or MICE-imputed training datasets (Table 5.4).

Due to the improvement in predictive performance observed in the holdout validation dataset, the missForest imputation method was initially pursued. However, the implementation of the preschool linear SVM on the missForest imputed training dataset demonstrated concerning performance patterns between the training and holdout sets, whereby the holdout performance was consistently higher compared to the training performance (Table 5.5) – an inversion of the train-test performance pattern expected of machine learning models. Further investigations identified that this train-test performance pattern fluctuated with different splits of the training and holdout validation datasets. Consequently, the MICE imputation method appeared to be a more robust imputation method as it generated similar predictive performance to the performance observed using the real training data in the complete data analysis. Moreover, the inverted train-test performance pattern was not observed. Therefore, the MICE imputation method was carried forward for final model development.

Table 5.4 Comparison of imputation methods using the preschool linear SVM on the complete preschool training dataset

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Complete data	0.88	0.67	0.86	0.39	0.96	0.63	0.91	10.26	0.63	0.48	302, 12, 31, 20
	0.85	0.66	0.78	0.40	0.92	0.45	0.91	5.27	0.65	0.43	146, 12, 15, 10
MICE	0. 90	0.70	0.88	0.43	0.97	0.71	0.91	15.05	0.59	0.54	305, 9, 29, 22
	0.85	0.66	0.77	0.40	0.92	0.43	0.91	4.86	0.65	0.42	145, 13, 15, 10
missForest	0.88	0.69	0.85	0.41	0.96	0.64	0.91	10.78	0.61	0.50	302, 12, 30, 21
	0.84	0.74	0.79	0.60	0.87	0.43	0.93	4.74	0.46	0.50	138, 20, 10, 15

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

Shaded rows report performance in the training set, unshaded rows report performance in the holdout validation set.

[a] The final column presents the confusion matrix for the model classifications, where TN=true negatives, FP=false positives, FN=false negatives, TP=true positives.

Table 5.5 Implementation of the imputation methods for the preschool linear SVM

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Initial Model	0.88	0.67	0.86	0.39	0.96	0.63	0.91	10.26	0.63	0.48	302, 12, 31, 20
	0.85	0.66	0.78	0.40	0.92	0.45	0.91	5.27	0.65	0.43	146, 12, 15, 10
MICE Imputation	0.87	0.64	0.82	0.31	0.97	0.64	0.89	10.32	0.71	0.42	979, 30, 122, 54
	0.85	0.66	0.81	0.40	0.92	0.45	0.91	5.27	0.65	0.43	146, 12, 15, 10
MissForest Imputation	0.86	0.62	0.75	0.27	0.97	0.59	0.88	8.17	0.76	0.37	976, 33, 129, 47
	0.85	0.66	0.79	0.40	0.92	0.45	0.91	5.27	0.65	0.43	146, 12, 15, 10

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

Shaded rows report performance in the training set, unshaded rows report performance in the holdout validation set.

[a] The final column presents the confusion matrix for the model classifications, where TN=true negatives, FP=false positives, FN=false negatives, TP=true positives.

5.3.4.2 Imbalanced data

Regardless of the degree of oversampling applied to the complete training dataset (Table A7), the discriminative ability of the preschool linear SVM remained constant (AUC: 0.77-0.79). Improvements in F_1-score, balanced accuracy, NPV and LR- were observed, however, the main benefit offered by oversampling was an improvement in model sensitivity (Figure 5.5, Table 5.6). Oversampling offered up to a 36% improvement in sensitivity compared to using the complete training dataset without oversampling. However, improvements in model sensitivity were only observed after oversampling the number of asthma cases by at least 200%. Specificity did reduce by up to 12% upon oversampling, yet this reduction was considered modest compared to the observed improvement in sensitivity.

As the main impact of oversampling was on the sensitivity of the preschool linear SVM, the effect of oversampling was similarly assessed on the preschool model that initially demonstrated the best sensitivity, the naïve Bayes model (Table 5.3). Interestingly, the improvement in sensitivity observed for the preschool linear SVM was not observed for the preschool naïve Bayes model – only a 4% improvement in sensitivity was observed across all degrees of oversampling. Thus, the impact of oversampling the training dataset on model performance appeared to be model dependent (Table A8, Figure A4).

Although oversampling increased the proportion of asthma cases within the training dataset, due to the low prevalence of asthma within the IOWBC, even a 300% increase in the number of asthma cases (which would result in 75% of asthma cases being synthetic data points) did not fully rectify the class imbalance (cases=204 and controls=314, ratio=2:3). Therefore, random undersampling of non-asthmatic controls was applied to completely balance the classes in the training dataset oversampled by 300% - the model trained on this dataset had demonstrated the best overall performance based on a combination of AUC and sensitivity (Table 5.6). Compared to the preschool linear SVM developed on the oversampled dataset, the model retrained on the oversampled and undersampled dataset (class ratio 1:1) offered improved performance in terms of both sensitivity (0.80 vs 0.76) and AUC (0.82 vs 0.78) (Figure 5.6, Table 5.7). For the other performance measures (apart from NPV), a modest reduction was observed.

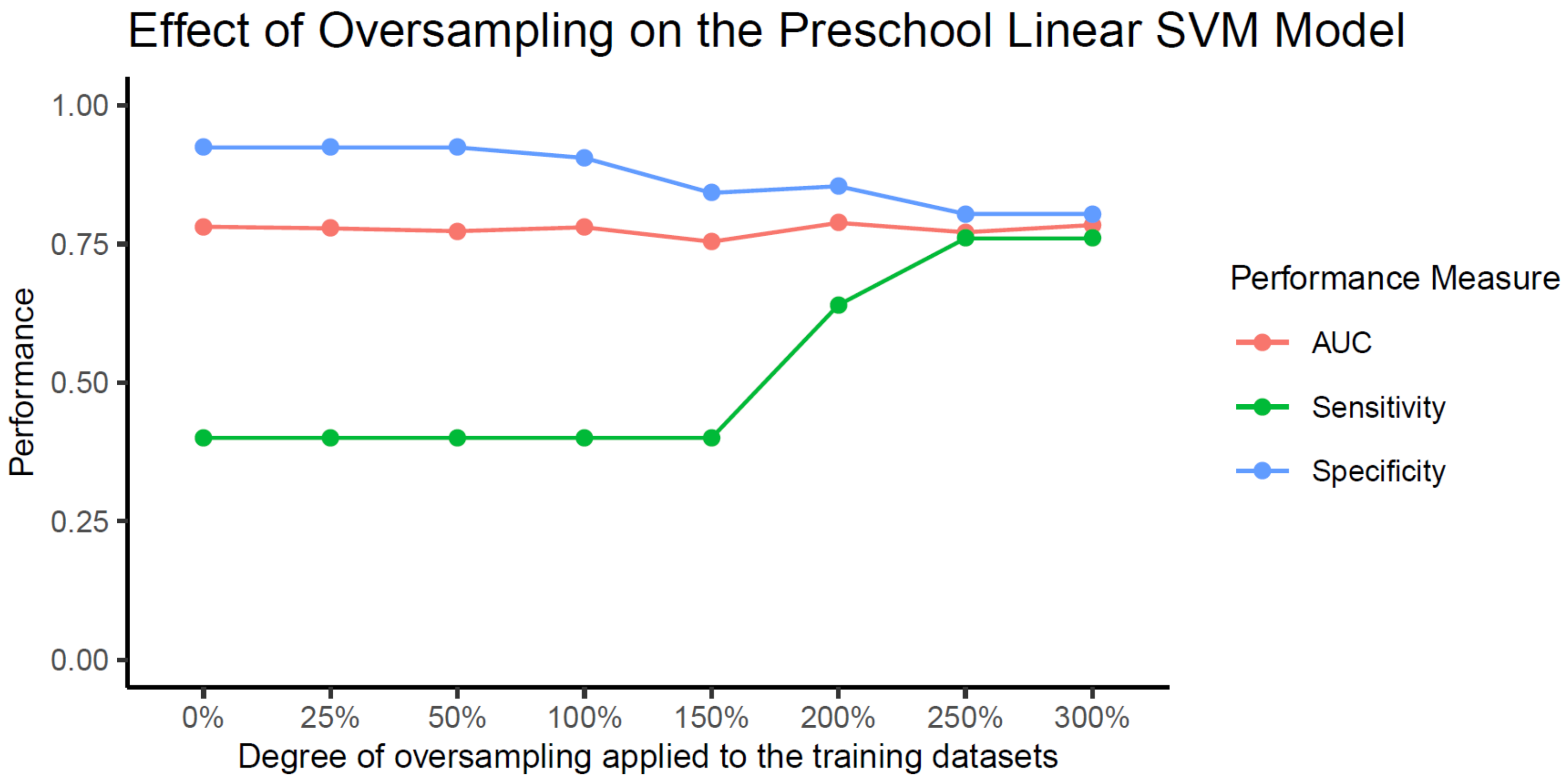


Figure 5.5 Effect of oversampling on the performance of the preschool linear SVM

Table 5.6 Performance of the preschool linear SVM upon the application of oversampling

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Initial Model	0.88	0.68	0.86	0.39	0.96	0.63	0.91	10.26	0.63	0.48	302, 12, 31, 20
	0.85	0.66	0.78	0.40	0.92	0.46	0.91	5.27	0.65	0.43	146, 12, 15, 10
Oversampled cases 25%	0.86	0.66	0.85	0.36	0.96	0.66	0.88	9.40	0.67	0.46	302, 12, 41, 23
	0.85	0.66	0.78	0.40	0.92	0.46	0.91	5.27	0.65	0.43	146, 12, 15, 10
Oversampled cases 50%	0.85	0.67	0.83	0.38	0.96	0.71	0.86	9.86	0.65	0.49	302, 12, 48, 29
	0.85	0.66	0.77	0.40	0.92	0.46	0.91	5.27	0.65	0.43	146, 12, 15, 10
Oversampled cases 100%	0.83	0.70	0.83	0.44	0.96	0.76	0.84	9.90	0.59	0.56	300, 14, 57, 45
	0.84	0.65	0.78	0.40	0.91	0.40	0.91	4.21	0.66	0.40	143, 15, 15, 10
Oversampled cases 150%	0.81	0.74	0.83	0.59	0.90	0.69	0.84	5.58	0.46	0.64	281, 33, 53, 75
	0.78	0.62	0.75	0.40	0.84	0.29	0.90	2.53	0.71	0.33	133, 25, 15, 10

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Oversampled cases 200%	0.90	0.76	0.84	0.63	0.88	0.72	0.83	5.36	0.42	0.67	277, 37, 57, 96
	0.83	0.75	0.79	0.64	0.85	0.41	0.94	4.40	0.42	0.50	135, 23, 9, 16
Oversampled cases 250%	0.80	0.78	0.86	0.72	0.84	0.72	0.84	4.49	0.34	0.72	264, 50, 51, 128
	0.80	0.78	0.77	0.76	0.80	0.38	0.96	3.87	0.30	0.51	127, 31, 6, 19
Oversampled cases 300%	0.79	0.78	0.84	0.71	0.84	0.74	0.82	4.46	0.34	0.72	264, 50, 59, 145
	0.80	0.78	0.78	0.76	0.80	0.38	0.96	3.87	0.30	0.51	127, 31, 6, 19

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

Shaded rows report performance in the training set, unshaded rows report performance in the holdout validation set.

[a] The final column presents the confusion matrix for the model classifications, where TN=true negatives, FP=false positives, FN=false negatives, TP=true positives.

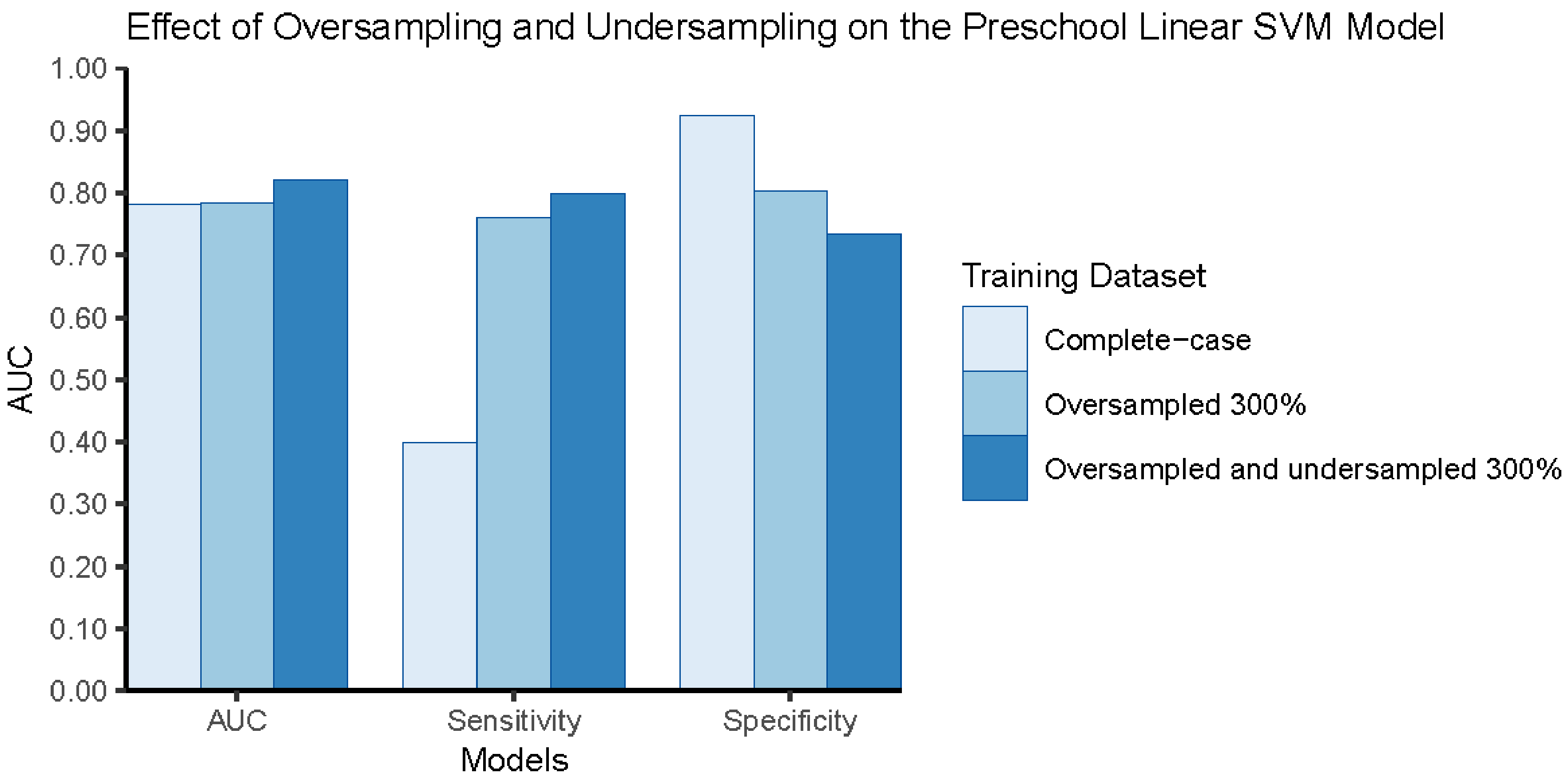


Figure 5.6 Effect of oversampling and undersampling on the preschool linear SVM

Table 5.7 Performance of the preschool linear SVM upon the application of oversampling and random undersampling

	Accuracy	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Initial Model	0.88	0.68	0.86	0.39	0.96	0.63	0.91	10.26	0.63	0.48	302, 12, 31, 20
	0.85	0.66	0.78	0.40	0.92	0.46	0.91	5.27	0.65	0.43	146, 12, 15, 10
Oversampled cases 300%	0.79	0.78	0.84	0.71	0.84	0.74	0.82	4.46	0.34	0.73	264, 50, 59, 145
	0.80	0.78	0.78	0.76	0.80	0.38	0.96	3.87	0.30	0.51	127, 31, 6, 19
Oversampled 300%, undersampled	0.78	0.78	0.85	0.80	0.77	0.78	0.79	3.47	0.26	0.79	157, 47, 41, 163
	0.74	0.77	0.82	0.80	0.73	0.32	0.96	3.01	0.27	0.46	116, 42, 5, 20

PPV: positive predictive value; NPV: negative predictive value; LR+: positive likelihood ratio; LR-: negative likelihood ratio; AUC=area under the curve.

Shaded rows report performance in the training set, unshaded rows report performance in the holdout validation set.

[a] The final column presents the confusion matrix for the model classifications, where TN=true negatives, FP=false positives, FN=false negatives, TP=true positives.

5.3.4.3 Implementation of optimisation strategies for final model development

The application of imputation, oversampling and random undersampling all demonstrated improvements in predictive performance compared to the preschool linear SVM initially developed using the complete training dataset. Therefore, as described in Chapter 5.2.5, all three training optimisation strategies were implemented for the selection of the final early life (CAPE) and preschool (CAPP) models. Due to computational constraints, and supported by its suboptimal performance compared to the other machine learning algorithms developed using the complete training datasets, the polynomial SVM was not carried forward for model redevelopment. In general, the preschool models which were developed offered improved performance (AUC ranged from 0.16-0.84) compared to the early life models (AUC ranged from 0.42-0.71) (further detail available at: https://doi.org/10.5258/SOTON/D1943).

5.3.5 Childhood Asthma Prediction in Early-life (CAPE) Model

The best performing early life model selected as the CAPE model was developed using an SVM classifier (RBF kernel, C=45.1 and gamma=0.005), trained on the complete early life training set, undersampled to balance class proportions (n=136, 68 asthmatics and 68 non-asthmatics). The model performed with an AUC=0.71 and Brier score=0.21 (Figure 5.7). Based on the threshold cut-off that maximised the Youden’s Index (threshold=0.42), classifications of asthma and no asthma were made and performance measures evaluated (Table 5.8). Based on this threshold, the CAPE model demonstrated moderate predictive power, with 71% accuracy balanced between the two classes, 74% sensitivity and 68% specificity in the early life holdout validation set. However, the model offered low PPV (26%), suggesting that a large number of false positive predictions were common.

5.3.5.1 External validation of the CAPE model

In MAAS, 1018 and 898 individuals had a defined asthma outcome at ages 8 and 11, respectively. The distribution of predictor data was largely similar between the IOWBC and MAAS (Table A9). However, a smaller proportion of individuals in MAAS were reported to have frequent wheeze in early life or at preschool age compared to in the IOWBC, likely due to discrepancies in the definitions used between the two cohorts. Individuals in MAAS were also more likely to have mothers with higher socioeconomic status compared to those in the IOWBC; this difference between cohorts may stem from individuals in the MAAS cohort being recruited from an affluent area[186].

To predict the development of asthma at the 8-year and 11-year time-points in MAAS, complete data on the eight CAPE predictors was available for 322 and 299 individuals, respectively. The CAPE model demonstrated moderate generalisability, maintaining an AUC of 0.71 at both 8 and 11 years in the unselected MAAS cohort (Table 5.8, Figure 5.7). Although the model showed good generalisability overall, up to a 7-9% reduction in PPV was observed at the same classification threshold evaluated in the IOWBC. Similarly, in the high-risk subgroups, despite a 3-4% increase in PPV, overall predictive performance decreased (Table 5.8).

5.3.6 Childhood Asthma Prediction at Preschool-age (CAPP) Model

The best performing classification algorithm selected as the CAPP model was an SVM classifier (linear kernel, C=0.33), trained on the complete preschool training dataset, with asthmatic cases oversampled by 300% and non-asthmatic controls undersampled to balance class proportions (n=408, 204 asthmatics, 204 non-asthmatics). This preschool model performed with an AUC of 0.82 and Brier score of 0.18 (Figure 5.7). Based on the threshold cut-off that maximised the Youden's Index (threshold=0.73), classifications of asthma and no asthma were made and performance measures evaluated (Table 5.9). Using this classification threshold, the CAPP model offered good predictive power, with 80% accuracy balanced between the two classes, 72% sensitivity and 88% specificity in the preschool holdout validation set. Compared to the CAPE model, the CAPP model also offered improved PPV (47%).

5.3.6.1 External validation of the CAPP model

For validation of the CAPP model in MAAS at the 8-year and 11-year time-points, complete data for the 12 CAPP predictors was available for 282 and 267 individuals, respectively. The model demonstrated good generalisability in predicting asthma at both 8 and 11 years (AUC=0.83 and 0.79, respectively) in the unselected MAAS subgroup (Table 5.9, Figure 5.7). PPV also remained comparable in MAAS (PPV=0.45 and 0.41, respectively), with further improvements reported in the high-risk subgroup validations at both time-points (Table 5.9).

Table 5.8 Performance of the CAPE model to predict school-age asthma in the IOWBC and MAAS

	Dataset	**Sample size (# asthmatic)**	**Balanced Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**LR+**	**LR-**	**F_1 Score**
IOWBC: 10 years	Training [a]	136 (68)	0.65	0.76	0.56	0.75	0.69	0.63	2.24	0.59	0.62
	Testing	255 (34)	0.71 (0.62-0.78)	0.71 (0.61-0.80)	0.74 (0.56-0.88)	0.68 (0.62-0.74)	0.26 (0.21-0.32)	0.94 (0.91-0.97)	2.29 (1.69-3.01)	0.39 (0.18-0.63)	0.38 (0.31-0.46)
MAAS: 8 years	Unselected	322 (38)	0.67 (0.60-0.74)	0.71 (0.63-0.79)	0.84 (0.71-0.95)	0.51 (0.45-0.56)	0.19 (0.16-0.21)	0.96 (0.93-0.99)	1.71 (1.40-2.03)	0.31 (0.10-0.57)	0.30 (0.26-0.35)
	Medium-risk	208 (31)	0.66 (0.59-0.73)	0.71 (0.61-0.80)	0.87 (0.74-0.97)	0.46 (0.39-0.53)	0.22 (0.19-0.25)	0.95 (0.91-0.99)	1.61 (1.31-1.95)	0.28 (0.06-0.59)	0.35 (0.30-0.40)
	High-risk	81 (16)	0.57 (0.45-0.67)	0.64 (0.47-0.80)	0.81 (0.63-1.00)	0.32 (0.22-0.43)	0.23 (0.18-0.28)	0.88 (0.75-1.00)	1.20 (0.86-1.56)	0.58 (0.00-1.35)	0.36 (0.27-0.43)
MAAS: 11 years	Unselected	299 (32)	0.68 (0.60-0.74)	0.71 (0.62-0.79)	0.84 (0.72-0.97)	0.51 (0.45-0.57)	0.17 (0.14-0.20)	0.96 (0.94-0.99)	1.72 (1.39-2.05)	0.31 (0.07-0.58)	0.28 (0.24-0.33)
	Medium-risk	192 (25)	0.67 (0.59-0.74)	0.71 (0.62-0.80)	0.88 (0.76-1.00)	0.47 (0.40-0.54)	0.20 (0.17-0.23)	0.96 (0.92-1.00)	1.65 (1.34-2.03)	0.26 (0.00-0.57)	0.32 (0.27-0.37)
	High-risk	72 (12)	0.58 (0.44-0.69)	0.60 (0.43-0.76)	0.83 (0.58-1.00)	0.33 (0.22-0.45)	0.20 (0.15-0.25)	0.91 (0.78-1.00)	1.25 (0.85-1.66)	0.50 (0.00-1.39)	0.32 (0.23-0.40)

[a] The CAPE model was developed using an SVM classification algorithm using a radial basis function kernel (C=45.1, gamma=0.0054), trained on the complete training dataset, with controls under-sampled to obtain a 1:1 class ratio.

Performance in the IOWBC validation and MAAS test sets are evaluated at thresholds of 0.42. In MAAS, performance was evaluated in the unselected population and among medium and high risk subgroups (defined as the child having at least one parent or both parents with allergic disease (asthma, eczema or allergic rhinitis), respectively.

Table 5.9 Performance of the CAPP models to predict school-age asthma in the IOWBC and MAAS

	Dataset	**Sample size (# asthmatic)**	**Balanced Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**LR+**	**LR-**	**F_1 Score**
IOWBC: 10 years	Training [a]	408 (204)	0.78	0.85	0.80	0.77	0.78	0.79	3.47	0.26	0.79
	Testing	183 (25)	0.80 (0.70-0.89)	0.82 (0.71-0.91)	0.72 (0.52-0.88)	0.88 (0.83-0.92)	0.47 (0.38-0.62)	0.95 (0.92-0.98)	5.99 (3.79-10.11)	0.32 (0.13-0.54)	0.56 (0.45-0.70)
MAAS: 8 years	Unselected	282 (33)	0.73 (0.64-0.81)	0.83 (0.75-0.90)	0.55 (0.36-0.70)	0.91 (0.88-0.95)	0.45 (0.33-0.59)	0.94 (0.92-0.96)	6.17 (3.64-10.69)	0.50 (0.33-0.69)	0.49 (0.36-0.62)
	Medium-risk	178 (26)	0.70 (0.60-0.80)	0.80 (0.70-0.88)	0.50 (0.31-0.69)	0.90 (0.85-0.95)	0.46 (0.32-0.63)	0.91 (0.89-0.94)	5.07 (2.77-9.95)	0.55 (0.34-0.77)	0.48 (0.32-0.63)
	High-risk	70 (13)	0.73 (0.59-0.87)	0.80 (0.62-0.94)	0.54 (0.23-0.77)	0.93 (0.86-0.98)	0.64 (0.40-0.90)	0.90 (0.84-0.95)	7.67 (2.92-39.46)	0.50 (0.24-0.81)	0.58 (0.32-0.78)
MAAS: 11 years	Unselected	267 (29)	0.73 (0.63-0.82)	0.79 (0.68-0.88)	0.55 (0.38-0.72)	0.90 (0.87-0.94)	0.41 (0.29-0.55)	0.94 (0.92-0.96)	5.71 (3.44-9.85)	0.50 (0.30-0.71)	0.47 (0.33-0.62)
	Medium-risk	169 (22)	0.72 (0.61-0.82)	0.76 (0.61-0.88)	0.55 (0.36-0.73)	0.89 (0.84-0.94)	0.43 (0.29-0.59)	0.93 (0.90-0.96)	5.01 (2.75-9.65)	0.51 (0.30-0.74)	0.48 (0.32-0.63)
	High-risk	64 (10)	0.75 (0.59-0.90)	0.73 (0.47-0.94)	0.60 (0.30-0.90)	0.91 (0.83-0.98)	0.55 (0.31-0.86)	0.92 (0.87-0.98)	6.48 (2.40-32.40)	0.44 (0.11-0.80)	0.57 (0.30-0.78)

[a] The CAPP model was developed using an SVM classification algorithm using a linear kernel (C=0.33), trained on the complete training dataset, with cases oversampled by 300% and controls under-sampled to obtain a 1:1 class ratio.

Performance in the IOWBC validation and MAAS test sets are evaluated at thresholds of 0.73. In MAAS, performance was evaluated in the unselected population and among medium and high risk subgroups (defined as the child having at least one parent or both parents with allergic disease (asthma, eczema or allergic rhinitis), respectively.

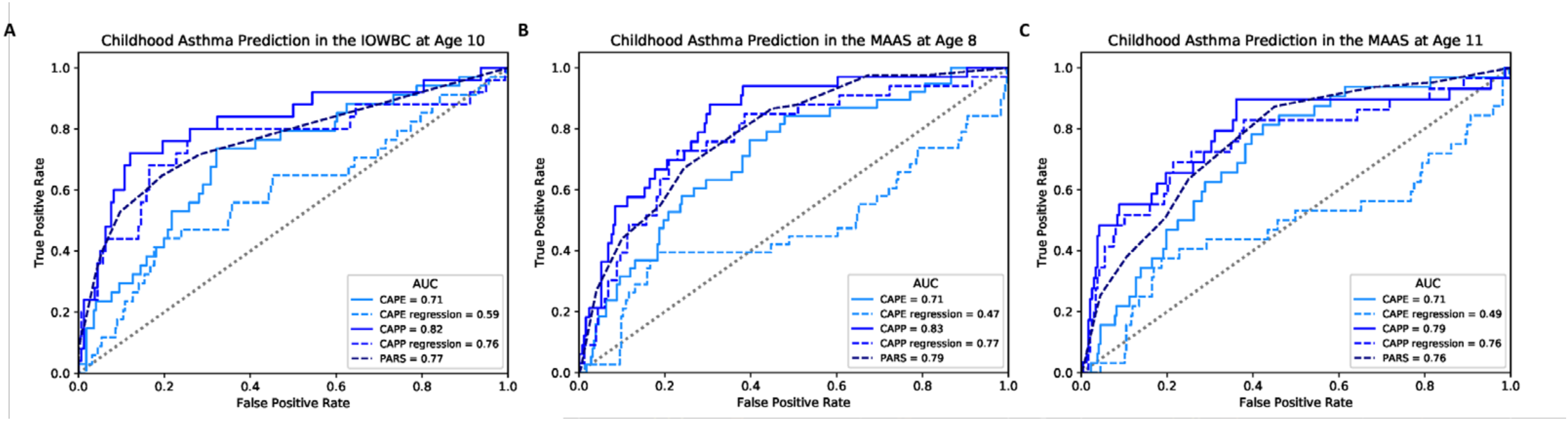


Figure 5.7 ROC curves comparing the performance of the CAPE and CAPP machine learning models, their equivalent regression models and PARS

Discriminative performance of each model is evaluated in the IOWBC holdout validation sets at age 10 (A) and upon validation in MAAS at age 8 years (B) and 11 years (C).

5.3.7 Sensitivity analysis

5.3.7.1 Predicting an alternative asthma definition

Asthma status, based on the alternative asthma definition incorporating BHR, was available for 1312 of the 1368 individuals analysed in the IOWBC (prevalence 8.61%). Despite an overall 92.3% agreement, there was a statistically significant difference between the two asthma definitions (p<0.01). This stemmed from a 97.6% agreement for labelling non-asthmatics but only a 53.8% agreement for labelling asthmatics (Figure 5.8).

A labelled asthma status using the alternative asthma definition was available for 248 out of the 255 individuals in the CAPE holdout validation dataset (20 asthmatic) and 179 out of the 183 individuals in the CAPP validation dataset (18 asthmatic). The CAPE and CAPP models were less robust to predict the alternative asthma outcome (CAPE AUC=0.67 vs 0.71 and CAPP AUC=0.79 vs 0.82). The CAPE and CAPP models were robust in correctly predicting non-asthmatics using the alternative asthma definition (similar NPV). However, neither model was robust in predicting asthmatics. Despite both models demonstrating an increased sensitivity to predict asthmatics, the corresponding increase in false positive predictions resulted in the PPV reducing by approximately 50% for both models, likely due to disagreement between the original and modified asthma definitions (Table 5.10, Figure 5.8).

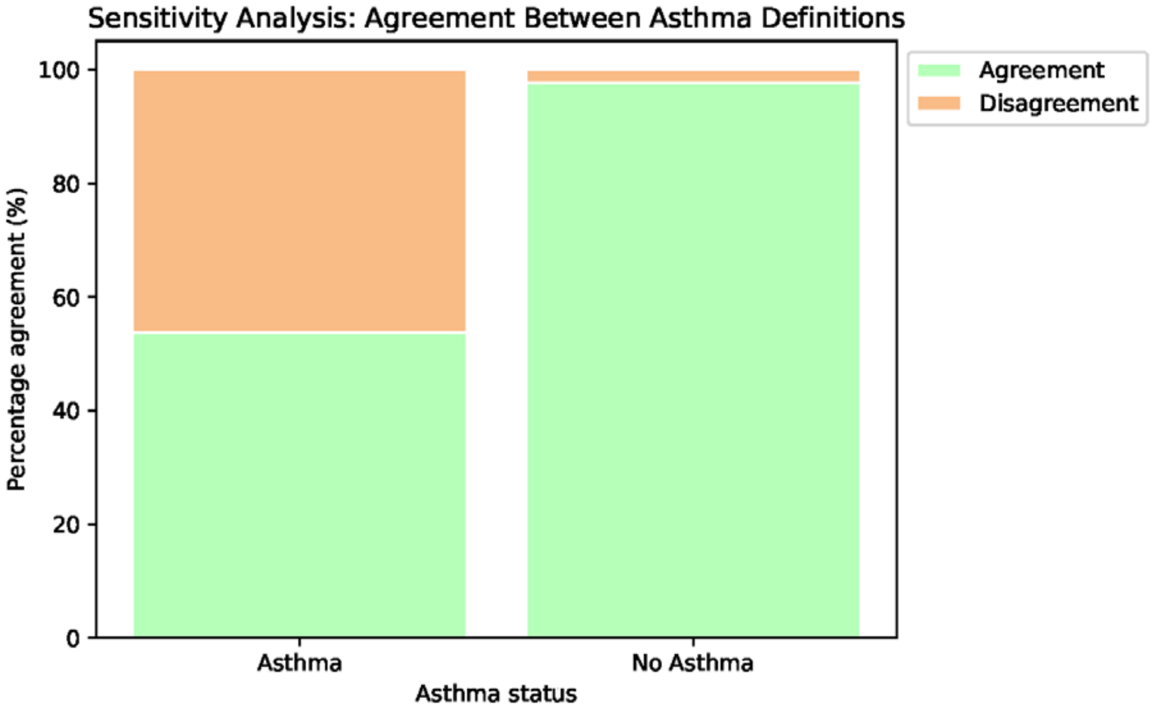


Figure 5.8 Agreement between the original and alternative asthma definitions

Of the 1368 individuals in the IOWBC included in the main study, 1312 individuals had their asthma status defined using the two asthma definitions: original definition used in the analysis (doctor diagnosis asthma ever and wheeze or use of asthma medication in the last 12 months) and an alternative definition (wheeze in the last 12 months and BHR). Each stacked bar represents the classification of individuals as asthmatic (left, n=160) or non-asthmatic (right, n=1152) based on the original asthma definition. Each bar shows the proportion of individuals for whom the alternative asthma definition assigned the same asthma status (green stacks) or opposing asthma status (orange stacks) compared to the original asthma definition.

Table 5.10 Performance of the CAPE and CAPP models to predict an alternative definition of school-age asthma

	Asthma definition [a] (% asthmatic)	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 score
CAPE	IOWBC Original (13.3%)	0.71 (0.62-0.78)	0.71 (0.61-0.80)	0.74 (0.56-0.88)	0.68 (0.62-0.74)	0.26 (0.21-0.32)	0.94 (0.91-0.97)	2.29 (1.69-3.01)	0.39 (0.18-0.63)	0.38 (0.31-0.46)
	IOWBC Alternative (8.1%)	0.70 (0.62-0.76)	0.67 (0.56-0.78)	0.90 (0.75-1.00)	0.49 (0.43-0.56)	0.13 (0.11-0.16)	0.98 (0.96-1.00)	1.77 (1.44-2.12)	0.20 (0.00-0.49)	0.23 (0.20-0.27)
	MAAS 8YR Alternative (5.0%)	0.61 (0.51-0.68)	0.69 (0.55-0.82)	0.87 (0.67-1.00)	0.34 (0.29-0.40)	0.07 (0.05-0.08)	0.98 (0.95-1.00)	1.32 (1.02-1.57)	0.39 (0.00-0.97)	0.12 (0.09-0.14)
	MAAS 11YR Alternative (3.0%)	0.60 (0.45-0.69)	0.58 (0.37-0.75)	0.86 (0.57-1.00)	0.33 (0.28-0.39)	0.03 (0.02-0.04)	0.99 (0.97-1.00)	1.29 (0.84-1.61)	0.43 (0.00-1.34)	0.06 (0.04-0.08)
CAPP	IOWBC Original (13.7%)	0.80 (0.70-0.89)	0.82 (0.71-0.91)	0.72 (0.52-0.88)	0.88 (0.83-0.92)	0.47 (0.38-0.62)	0.95 (0.92-0.98)	5.99 (3.79-10.11)	0.32 (0.13-0.54)	0.56 (0.45-0.70)
	IOWBC Alternative (10.1%)	0.77 (0.68-0.85)	0.79 (0.67-0.89)	0.83 (0.67-1.00)	0.71 (0.64-0.78)	0.25 (0.19-0.31)	0.97 (0.95-1.00)	2.92 (2.11-4.07)	0.23 (0.00-0.48)	0.38 (0.30-0.46)
	MAAS 8YR Alternative (5.3%)	0.68 (0.56-0.78)	0.70 (0.57-0.82)	0.79 (0.57-1.00)	0.57 (0.51-0.63)	0.09 (0.07-0.12)	0.98 (0.96-1.00)	1.83 (1.29-2.39)	0.38 (0.00-0.77)	0.17 (0.12-0.21
	MAAS 11YR Alternative (2.4%)	0.71 (0.53-0.81)	0.68 (0.40-0.87)	0.83 (0.50-1.00)	0.58 (0.52-0.64)	0.05 (0.03-0.06)	0.99 (0.98-1.00)	1.98 (1.15-2.63)	0.29 (0.00-0.88)	0.09 0.05-0.12)

[a] The outcome of school-age asthma was defined as follows: original asthma definition= doctor diagnosis of asthma ever plus the presence of wheeze or use of asthma medication in the last 12 months; alternative asthma definition= current wheeze and bronchial hyper-responsiveness. Both asthma outcomes were evaluated at age 10 in the IOWBC among in individuals in the respective validation sets for each model.

5.3.7.2 Predicting wheeze phenotypes

In the CAPE and CAPP holdout validation datasets, 213 and 167 individuals had a defined wheeze phenotype, respectively. Both models showed excellent power to predict the persistent wheeze phenotype, with 100% and 90% of cases correctly identified by the CAPE and CAPP models, respectively (Figure 5.9). For external validation in MAAS, 237 and 216 individuals with complete predictor and school-age asthma data for the CAPE and CAPP models, respectively, also had a defined wheeze phenotype. Among these individuals, the CAPE and CAPP models were able to offer a positive prediction to 90% and 57% of individuals with a persistent wheeze phenotype (Figure 5.9).

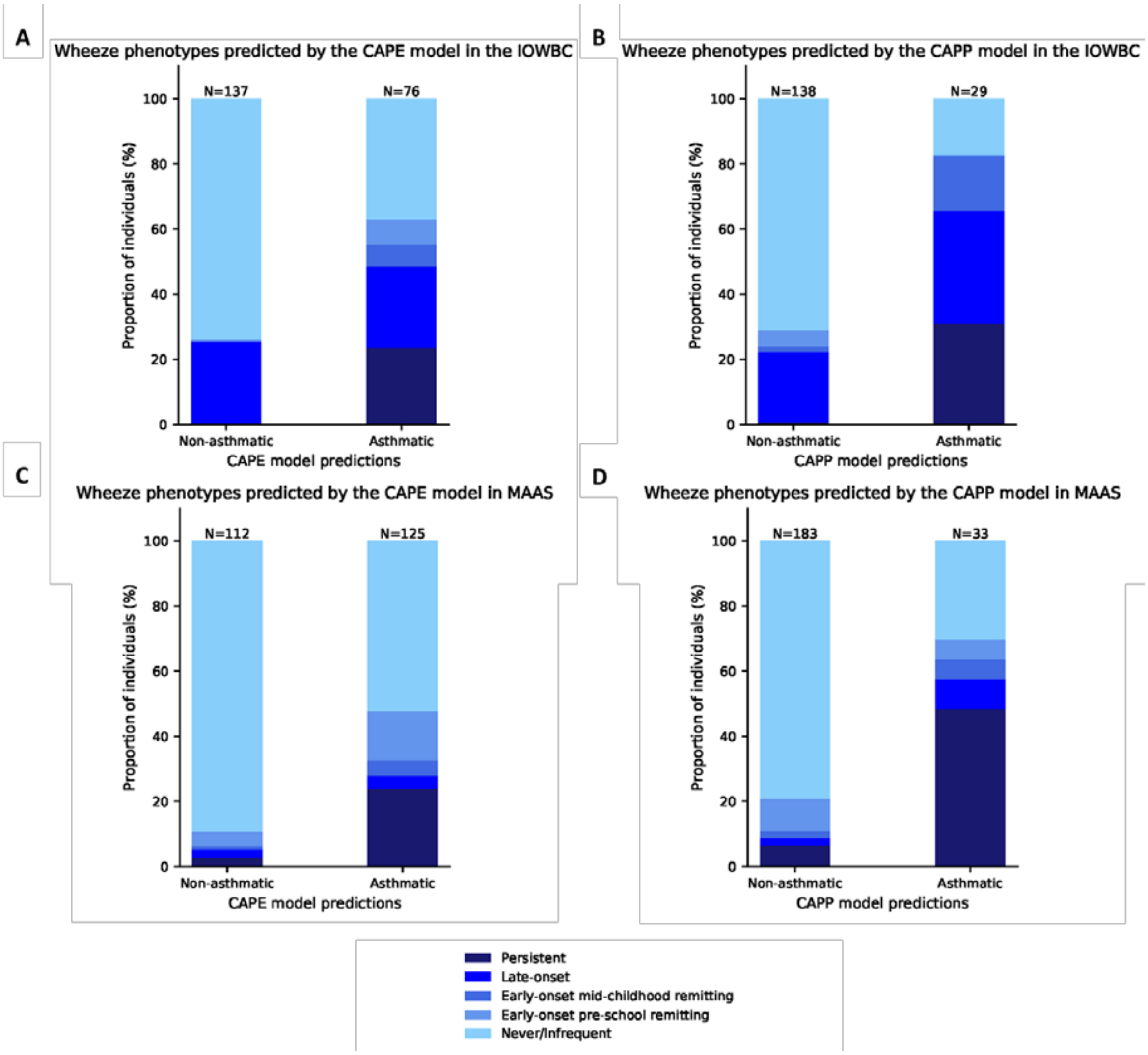


Figure 5.9 Predictions of wheeze phenotypes using the CAPE and CAPP models

The proportion of individuals assigned to each wheeze phenotype is presented for those offered a negative (non-asthmatic) or positive (asthmatic) prediction by either the CAPE (A) or CAPP model (B) in the IOWBC holdout validation datasets. Results were externally validated in MAAS for both the CAPE (C) and CAPP (D) models.

5.3.7.3 Comparison with regression methods

Both the CAPE and CAPP models outperformed their equivalent logistic regression models (Table 5.11, Figure 5.7). There was a substantial decline in predictive performance of the CAPE-logistic regression model (AUC=0.71 to 0.59), with predictions being no better than chance in MAAS at 8 and 11 years (AUC=0.47 and 0.49, respectively). Predictive power of the CAPP-logistic regression model was also lower compared to the CAPP-machine learning model (AUC=0.82 to 0.76, PPV=0.47 to 0.33).

Whilst the benchmark regression-based model for the CAPE model (Persistent Asthma Predictive Score, PAPS)[232] was unable to be replicated due to lack of data on key predictors in the IOWBC, the model comparable with the CAPP model, PARS (Paediatric Asthma Risk Score)[234], was replicated in the IOWBC and MAAS. In line with the replication of PARS conducted in Chapter 4, all individuals with complete data for the PARS predictors and the asthma outcome were included in the analysis (predicting asthma in the IOWBC at age 10: n=913, in MAAS at age 8 years: n=552, in MAAS at age 11 years: n=487). PARS demonstrated good predictive power in both the IOWBC and MAAS (AUC IOWBC=0.77, MAAS 8YR=0.79, MAAS 11YR=0.76) (Figure 5.7). Positive net reclassification indices among individuals with predictions available for both the CAPP and PARS models indicate that reclassifications made by the CAPP model offered equal, if not greater, accuracy to predict future asthmatics than the PARS model in both the IOWBC (Table 5.12) and MAAS (Table 5.13). For example, 32% of true asthmatic individuals who were incorrectly predicted as non-asthmatic by PARS were correctly reclassified by the CAPP model in the IOWBC.

Table 5.11 Comparison of the CAPE and CAPP models developed using machine learning and traditional logistic regression algorithms

	Algorithm (dataset)	**BA**	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**LR+**	**LR-**	**F$_1$ score**
CAPE	SVM [a] (IOWBC test set)	0.71 (0.62-0.78)	0.71 (0.61-0.80)	0.74 (0.56-0.88)	0.68 (0.62-0.74)	0.26 (0.21-0.32)	0.94 (0.91-0.97)	2.29 (1.69-3.01)	0.39 (0.18-0.63)	0.38 (0.31-0.46)
	Logistic regression [b] (IOWBC test set)	0.62 (0.54-0.71)	0.59 (0.48-0.70)	0.44 (0.27-0.59)	0.80 (0.75-0.85)	0.25 (0.17-0.35)	0.90 (0.88-0.93)	2.22 (1.33-3.43)	0.70 (0.49-0.91)	0.32 (0.21-0.43)
	Logistic regression (MAAS 8YR)	0.60 (0.52-0.68)	0.47 (0.36-0.59)	0.39 (0.24-0.55)	0.80 (0.75-0.85)	0.21 (0.13-0.29)	0.91 (0.89-0.93)	2.00 (1.15-3.02)	0.75 (0.55-0.96)	0.28 (0.17-0.38)
	Logistic regression (MAAS 8YR)	0.58 (0.50-0.67)	0.49 (0.36-0.61)	0.38 (0.22-0.53)	0.79 (0.74-0.84)	0.18 (0.10-0.25)	0.91 (0.89-0.94)	1.79 (0.96-2.84)	0.79 (0.57-1.01)	0.24 (0.14-0.34)
CAPP	SVM [c]	0.80 (0.70-0.89)	0.82 (0.71-0.91)	0.72 (0.52-0.88)	0.88 (0.83-0.92)	0.47 (0.38-0.62)	0.95 (0.92-0.98)	5.99 (3.79-10.11)	0.32 (0.13-0.54)	0.56 (0.45-0.70)
	Logistic regression [d] (IOWBC test set)	0.77 (0.68-0.85)	0.76 (0.63-0.88)	0.80 (0.64-0.96)	0.74 (0.67-0.80)	0.33 (0.26-0.41)	0.96 (0.93-0.99)	3.08 (2.24-4.34)	0.27 (0.05-0.50)	0.47 (0.38-0.56)
	Logistic regression (MAAS 8YR)	0.72 (0.64-0.78)	0.77 (0.67-0.85)	0.82 (0.70-0.94)	0.61 (0.55-0.67)	0.22 (0.18-0.26)	0.96 (0.94-0.99)	2.12 (1.68-2.66)	0.30 (0.10-0.52)	0.35 (0.29-0.41)
	Logistic regression (MAAS 8YR)	0.71 (0.62-0.78)	0.76 (0.64-0.86)	0.79 (0.66-0.93)	0.62 (0.56-0.68)	0.20 (0.17-0.25)	0.96 (0.93-0.99)	2.10 (1.62-2.67)	0.33 (0.11-0.59)	0.32 (0.26-0.39)

The CAPE and CAPP machine learning and equivalent logistic regression models were evaluated at age 10 in the IOWBC, in individuals in the respective holdout validations sets for each model. Validation in MAAS was performed to evaluate the prediction of asthma at 8 years (MAAS 8YR) and 11 years (MAAS 11YR). Performance measures were evaluated at the optimal model thresholds, which maximised the Youden's Index: [a]=0.42, [b]=0.48, [c]=0.73, [d]=0.42.

Table 5.12 Reclassification table comparing changes in prediction categorisation between the PARS and CAPP models in the IOWBC

	Predicted risk (CAPP model)				**Reclassified by CAPP (%)**			
Predicted risk (PARS model)	**No asthma**	**Asthma**	**Total**		**Increased risk**	**Decreased risk**	**Correctly reclassified**	**NRI**
No asthma at age 10 (n=149)								
No asthma	130	**9** [b]	139					
Asthma	**1** [a]	9	10		9(6%)	1(<1%)	1(<1%)	-0.05
Total	131	18	149					
Asthma at age 10 (n=25)								
No asthma	7	**8** [a]	15					
Asthma	**0** [b]	10	10		8(32%)	0(0%)	8(32%)	0.32
Total	7	18	25					
				Total	**17**	**1**	**9**	

Reclassification table comparing the change in individual asthma predictions with the CAPP model instead of the PARS model (reference model). For the PARS model, categorisations of predictions as asthmatic and non-asthmatic were based on the optimal threshold (cut-off=7) as defined in the original publication[234]. Results are presented separately for individuals who were asthmatic and non-asthmatic at age 10. Values in bold identify the number of individuals who were reclassified into a more appropriate ([a]) or less appropriate ([b]) risk group by the CAPP model with respect to the risk classifications made by the PARS model. NRI=net reclassification index is given separately for true asthmatics and non-asthmatics.

Table 5.13 Reclassification table comparing changes in prediction categorisation between the PARS and CAPP models in MAAS

		Predicted risk (CAPP model)			**Reclassified by CAPP (%)**			
	Predicted risk (PARS model)	**No asthma**	**Asthma**	**Total**	**Increased risk**	**Decreased risk**	**Correctly reclassified**	**NRI**
MAAS 8YR (PARS AUC=0.86 vs CAPP=0.83)								
No asthma (n=213)	**No asthma**	173	**14** [b]	187				
	Asthma	**21** [a]	5	26	14(7%)	21(10%)	21(10%)	0.03
	Total	194	19	213				
Asthma (n=28)	**No asthma**	5	**7** [a]	12				
	Asthma	**7** [b]	9	16	7(25%)	7(25%)	7(25%)	0.00
	Total	12	16	28	21	28	28	
MAAS 11YR (PARS AUC=0.78 vs CAPP=0.79)								
No asthma (n=215)	**No asthma**	170	**14** [b]	184				
	Asthma	**24** [a]	7	31	14(7%)	24(11%)	24(11%)	0.05
	Total	194	21	215				
Asthma (n=26)	**No asthma**	8	**7** [a]	15				
	Asthma	**4** [b]	7	11	7(27%)	4(15%)	7(27%)	0.12
	Total	12	14	26	21	28	21	

Reclassification table comparing the change in individual asthma predictions with the CAPP model instead of the PARS model (reference model). For the PARS model, categorisations of predictions as asthmatic and non-asthmatic were based on the optimal threshold (cutoff=7) as defined in the original publication[234]. Results are presented separately for individuals who were asthmatic and non-asthmatic at ages 8 and 11. Values in bold identify the number of individuals who were reclassified into a more appropriate ([a]) or less appropriate ([b]) risk group by the CAPP model with respect to the risk classifications made by the PARS model. NRI=net reclassification index is given separately for true asthmatics and non-asthmatics.

5.3.8 Explaining the “black-box” models

Based on the SHAP values generated for each model, only a subset of predictors were shown to have a major contribution on the predictions – early life cough and wheeze for the CAPE model and preschool cough, atopy and polysensitisation for the CAPP model (Figure 5.10). The contributions of these predictors over others in the models were reinforced upon the extraction of local explanations for individual predictions (examples presented in Figure 5.11). Redevelopment of the CAPE and CAPP models, including only those predictors shown to offer major contributions to the classifications, showed similar performance for the CAPP model but a decline in performance for the CAPE model (10% fall in AUC) (Figure 5.12, Table A10, Table A 11).

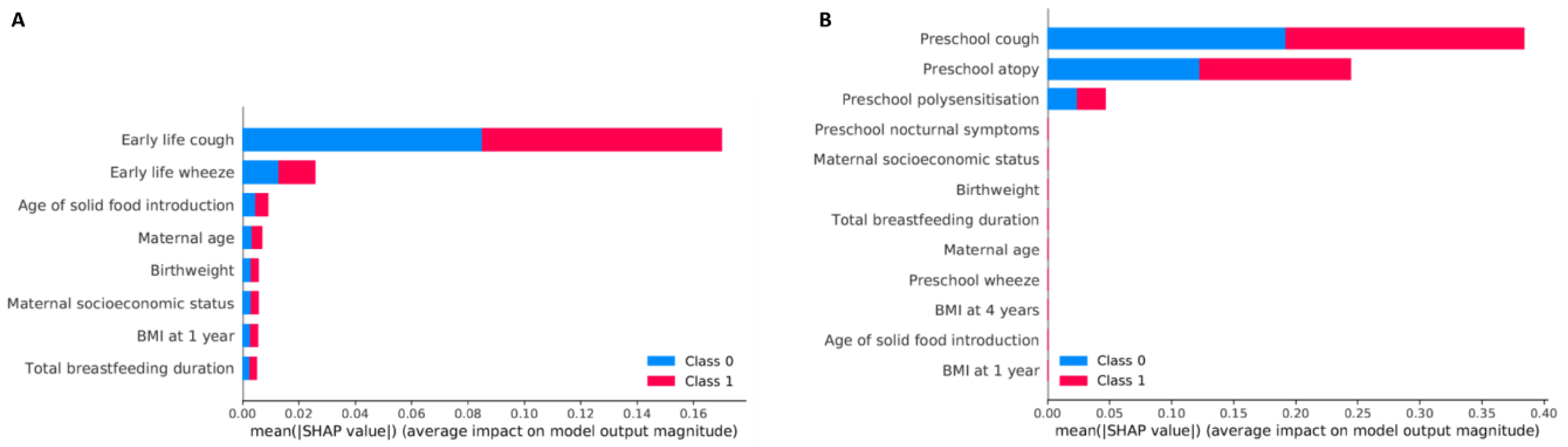


Figure 5.10 Global interpretations of the CAPE and CAPP models based on SHAP

The stacked bar plot shows the mean absolute SHAP value of each feature across all samples in the IOWBC validation sets for the CAPE (A) and CAPP (B) models. The bars show the impact each feature has on the model offering a prediction of asthma (Class 1 – red bars) and no asthma (Class 0 – blue bars). The order of the bars corresponds to the contribution (feature importance) each predictor makes in determining the output of the models (top to bottom = highest to lowest contribution).

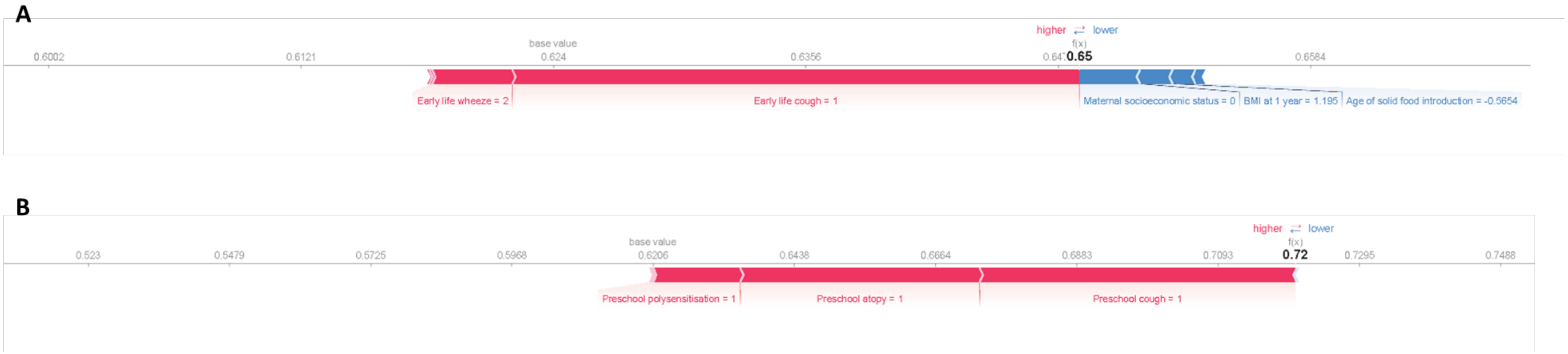


Figure 5.11 Local interpretation of individual predictions made by the CAPE and CAPP models based on SHAP

Example explanations of the predictions offered by the CAPE (A) and CAPP (B) models for two randomly selected individuals in the IOWBC validation sets are shown. (A) The individual was offered a predicted probability (f(x)) of 0.65 for developing school-age asthma at age 10. The plot shows that frequent early life wheeze and early life cough positively contributed to increase the probability of the individual being classified as asthmatic. Low maternal socioeconomic status, higher than average BMI and earlier than average introduction of solid foods into the diet negatively impacted a prediction of asthma, contributing to a reduction in the overall predicted probability of asthma. (B) The individual was offered a predicted probability of 0.72 for developing school-age asthma by the CAPP model. Having preschool cough and being both atopic and polysensitised at age 4 contributed to an increase in the predicted probability of asthma. The base value is the reference point the feature contributions take effect from and is defined as the prediction that would be made for an asthmatic individual if none of the features of the current output were known.

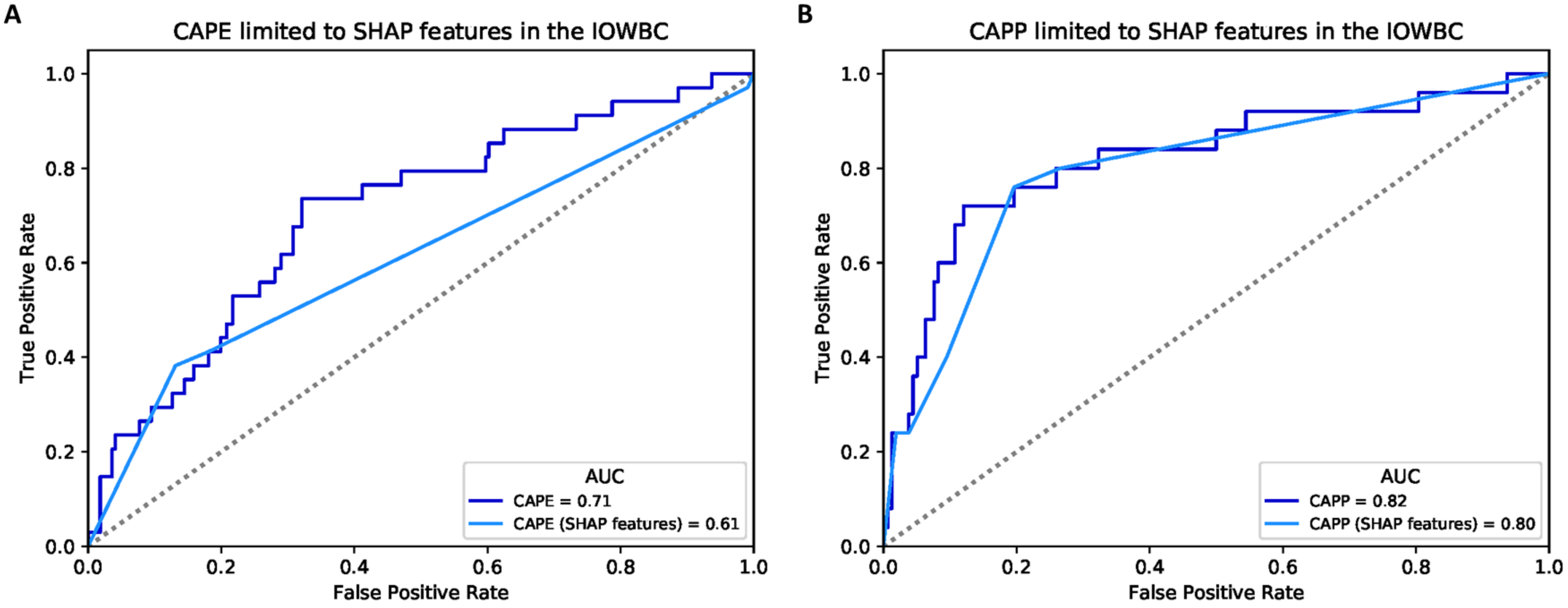


Figure 5.12 ROC curves comparing the performance of the original and SHAP feature restricted CAPE and CAPP models

Performance is presented for the CAPE (A) and CAPP (B) models in the IOWBC using all features selected through RFE (dark blue line) and the subset of predictors identified as offering the major contributions to model predictions (light blue line) – CAPE= early life cough and wheeze, CAPP= preschool cough, atopy and polysensitisation.

5.4 Discussion

5.4.1 Summary of findings

Two models, predicting school-age asthma at age 10 within the general population of the IOWBC, were developed using machine learning classification methods. The CAPE model uses an RBF SVM classifier and eight predictors to predict school-age asthma in early life. The CAPP model uses a linear SVM classifier and twelve predictors available within the first four years of life. Both machine learning models offered superior predictive power and generalisability upon external validation compared to equivalent models developed using logistic regression methods as well as existing regression-based models. Whilst the primary prediction outcome was school-age asthma, both models demonstrated greater sensitivity to predict individuals likely to experience persistent wheeze throughout childhood.

5.4.2 Comparisons with existing models

To date, twenty-one regression-based prediction models have been developed for childhood asthma (reviewed in Kothalawala *et al.*[109] and detailed in Chapter 3), of which only six have been externally validated. A recent systematic review further identified 10 studies that developed prediction models for childhood asthma using machine learning approaches, but only eight specifically predicted school-age asthma (5-14 years)[250]. Another study published after the review directly compared the performance of a current regression-based asthma prediction model (PARS) with a conditional inference tree-based decision rule model using the same predictors[251]. However, none of these studies have externally validated the machine learning models which they have proposed.

Similar to the CAPE and CAPP models, most published asthma prediction models are very good at ruling out asthma rather than ruling in asthma. This may be due to asthma being inherently highly heterogeneous, with the non-specific and transient nature of symptoms used to define asthma resulting in frequent misdiagnoses. However, the difficulty for many existing prediction models to rule in asthma is also likely a consequence of low study power due to low asthma prevalence[109]. Even if existing models offer good PPV, this often degrades upon validation. Indeed, despite having similar asthma prevalence to existing studies in the original training set, the machine learning-based CAPP model offered a 30% improvement in sensitivity compared to the best regression model described to date (sensitivity: CAPP=0.72, loose API=0.42)[219] and further 10% improvement in PPV compared to its benchmark model, PARS. This is consistent with Owora *et al*.'s novel tree-based model offering better predictive performance compared to an equivalent regression-based PARS model (AUC=0.85 vs 0.71)[251]. Many of the other machine learning models also demonstrated greater performance to predict asthma than existing regression-based models[250]. However, with low sample sizes and indications of overfitting in many of these studies, the lack of external validation renders it impossible to evaluate any superior performance offered by these models, especially since they were all developed in high-risk populations.

Whilst improvements in the CAPE and CAPP models over comparable regression-based models may stem from the use of more complex machine learning algorithms, it is important to acknowledge that the improved performance may be a result of the optimised training techniques that were applied to train and develop the CAPE and CAPP models on balanced datasets. Existing models have often been developed using highly imbalanced datasets and failed to address the imbalance using resampling methods. The analyses conducted in Chapter 5.3.4.2 highlighted that for the same model, the application of oversampling and undersampling has potential to improve predictive performance. However, this was also demonstrated to be model dependent. Of the asthma prediction models identified to date, only one study addressed the issue of class imbalance; in that study, Bose *et al*. compared a number of undersampling methods but observed only marginal improvements in the performance of their machine learning model in their dataset[258]. As class balance techniques have not been applied to studies that have developed asthma prediction models using logistic regression methods, it is difficult to ascertain whether the improved performance of the CAPE and CAPP models is due to the use of more sophisticated methods or merely the use of a dataset that offers a better balance of training opportunities to predict each outcome class.

Within this study, the comparison of the machine learning models with equivalent logistic regression models aimed to account for differences in the training dataset and directly compare the algorithms used to develop the models. As the CAPP machine learning model was more generalisable and retained its positive predictive power upon replication compared to its equivalent logistic regression model, it supports the hypothesis that the use of more complex machine learning algorithms may be able to address limitations of logistic regression models that often struggle to identify true asthmatics and present with modest generalisability.

Furthermore, reclassification tables comparing the CAPP and PARS models were suggestive of the CAPP model being able to predict future asthmatics better than PARS, with a greater proportion of correct reclassifications than incorrect reclassification made by the CAPP model in both IOWBC and MAAS. However, this needs to be confirmed within a larger cohort. The moderate but limited predictive power of the CAPE model compared to the CAPP model was unsurprising given the known difficulty of predicting the future development of childhood asthma in the first few years of life[2]. Yet, using machine learning approaches, the CAPE model was also able to offer improved discriminative performance compared to its existing regression-based benchmark model.

5.4.3 Predictor selection and availability

Both the CAPE and CAPP models include data collected across multiple time-points. Given the variable nature of asthma development and risk throughout early childhood, the consideration of predictors across multiple time-points allowed identification of a novel combination of predictors that together improved the ability of the models to predict asthma. Whilst data collected across multiple time-points may hinder the utility of the prediction models, the selected predictors are all typically reported during routine health visits or tracked in child health records. Only the predictors of atopy and polysensitisation, which require a SPT, may restrict the applicability of the CAPP model in primary care. However, these predictors are well-established in the literature, were shown to make large contributions to the predictions (Figure 5.10), and resulted in a 10% reduction in AUC when excluded from the model (Table A12). Hence, the predictive benefit offered by the inclusion of sensitisation was deemed to outweigh the potential reduction in applicability.

Of the predictors selected for inclusion in the two models, some were well-established risk factors with a clear inferred direction of asthma risk (Figure 5.3). Others were predictors which have not previously been used in asthma prognostic prediction models and offer a less clear direction of asthma risk (maternal age at the time of the child’s birth, age of solid food introduction and total breastfeeding duration). The selection of these newly selected predictors, over others that have more established biological relevance in the literature (such as parental asthma, eczema or allergic rhinitis), may be cautiously accepted by the clinical community. However, RFE identifies the subset of features that collectively offer the best predictive accuracy rather than devising a comprehensive list of childhood asthma risk factors, which may be biologically sound but lacking in predictive power[190]. In fact, the predictors of wheeze and cough were among those repeatedly included in the majority of machine learning models identified to date[250]. The predictors of atopy, polysensitisation and wheeze were also included in Owora *et al.*’s machine learning model, however, in this model, the predictors were taken from the PARS model rather than being identified from an independent feature selection[251]. It is also important to acknowledge the possibility that the selection of these newly selected predictors may stem from an inherent bias of the random forest algorithm to assign greater importance to features which are continuous or which have a large number of categories[260]. However, as the CAPE and CAPP models developed using these selected predictors demonstrated improved performance against existing prediction models, any bias stemming from the feature selection process did not appear to limit the inclusion of features that were truly predictive of school-age asthma.

5.4.4 Prediction generalisability, robustness and resolution

In the unselected MAAS cohort, the CAPE and CAPP models showed similar performance to predict asthma across school ages as observed in the IOWBC (despite the marginal decline in the PPV of the CAPE model). Validation in medium and high-risk MAAS subgroups showed the PPV of both models to increase with the number of allergic parents, suggesting that confidence in ruling in asthma improves in high-risk groups; but replication in a larger study population is required.

The lack of a clear definition for asthma is an unavoidable limitation in epidemiological studies[249]. The asthma definition used in this study aimed to account for children with a clinical indication of asthma (physician diagnosed) who were actively symptomatic, but also those potentially asymptomatic at the time of assessment due to the use of symptom-relieving medications. Whilst both models were robust in predicting non-asthmatics using the alternative asthma definition of wheeze and bronchial hyper-responsiveness (BHR), they had reduced power to predict true asthmatics (~50% decline in PPV). The latter may be explained by objective tests, such as spirometry and BHR, being influenced by treatment; potential asthmatics on controller medications, whom the models are trained to identify as asthmatic, may be considered as non-asthmatic with the alternative definition, resulting in greater false positive predictions. It is also possible that the arbitrary nature of the discrete PC_{20} cut-off used confirm a positive BHR could have censored the resulting asthma diagnoses. For example, upon challenge, some individuals may have experienced a substantial decline in FEV_1, but if their FEV_1 failed to decline past 20%, it would not have been possible to calculate the PC_{20} needed to identify BHR. This limitation in the alternative asthma definition may have been avoided if a continuous measure of BHR, such as that obtained using a dose-response slope, was used.

As the aim of this study was to compare whether machine learning approaches could improve upon existing regression-based models that predict childhood asthma, the primary prediction outcome for this study was restricted to school-age asthma rather than predicting asthma phenotypes. However, acknowledging the importance of exploring specific sub-phenotypes of asthma, the resolution of the machine learning models to predict an individual's childhood wheeze trajectory was explored. Notably, both the CAPE and CAPP models showed excellent sensitivity to predict individuals with a persistent wheeze phenotype; these individuals would likely benefit from early primary or secondary asthma prevention/ management.

To promote the clinical use of complex machine learning methods, studies must address the major hurdle of model interpretability. Studies such as Bose *et al*.'s have attempted to address the issue of model interpretability using feature importance measures that generate an importance ranking for the predictors included in the model. However, feature importance is limited in that it is unable to offer insight into the direction of the predictor's effect or provide information on how predictors interact to deduce individual predictions. By using SHAP, such information was extractable from the CAPE and CAPP "black-box" machine learning models, and enabled both global and local explanations of model predictions to be uncovered.

5.4.5 Strengths and limitations

This study had a number of strengths. First, each model was developed to make timely predictions to identify future asthmatics within a general population, rather than among those already considered at high-risk (mainly those experiencing wheeze or with a familial history of asthma/allergy). Second, by utilising machine learning methods, new predictors of school-age asthma were selected, and the models which were subsequently developed offered improved predictive performance over current regression-based methods. Third, to our knowledge, this is the first study to externally validate childhood asthma prediction models developed using machine learning approaches. The models demonstrated good generalisability to predict school-age asthma across multiple time-points, without degrading the predictive power to rule in asthma (particularly with the CAPP model). Fourth, the two models displayed excellent sensitivity to predict a subgroup of individuals with persistent wheeze. Finally, this study used SHAP to address one of the key issues preventing the uptake of machine learning methods in clinical practice - the inability to interpret the models and explain the individual predictions made[261].

However, this study was limited by both model development and validation being conducted in UK birth cohorts with predominantly Caucasian populations, potentially limiting generalisability among populations from different genetic and environmental backgrounds. Machine learning also requires large datasets – the introduction of more data would undoubtedly improve the performance of the developed machine learning models further. To retain a sample size appropriate for machine learning, feature selection was conducted before performing a train-test split. This decision could have resulted in information leakage, potentially biasing the performance seen in the IOWBC holdout validation sets. To mitigate any bias, external replication was performed to evaluate the models– as performance in MAAS was similar to that observed in the IOWBC, data leakage was not deemed a significant problem. A valid alternative approach would have been to perform a nested cross-validation, enabling all the data to be used for training and testing whilst simultaneously obtaining confidence intervals to assess the generalisability of the predictions. In addition, as there is no gold standard threshold for asthma prediction models, performance measures for the CAPE and CAPP models were evaluated at the classification threshold that maximised the Youden's index. Whilst this is in line with methods used among current studies, the threshold cut-off used for classification will impact the performance being reported. Hence, until a consensus is reached within the clinical community on the most appropriate threshold to use, performance measures stemming from the confusion matrix (detailed in Chapter 2.3.6) must be evaluated with an appreciation that different thresholds have been used between studies. Finally, whilst genomic data was available in the IOWBC, only clinical, environmental and simple biomarker (such as SPTs) predictors were considered. It is possible that the consideration of genomic predictors might significantly improve childhood asthma predictions further[222,256]; however, the aim of this study was to explore whether machine learning methods could surpass the predictive ceiling that existing logistic regression methods appeared to be limited to. Hence, to provide a fair comparison with existing regression-based models, asthma genomic biomarkers were not incorporated at this stage of the study.

5.4.6 Conclusion

Using machine learning, the CAPE and CAPP models were able to surpass the predictive performance of similar models developed using traditional regression-based methods. Both models were generalisable in an independent population, with the CAPP model also demonstrating superior predictive power to rule in true asthmatics compared to its benchmark model (and was retained upon validation). Both models also demonstrated excellent sensitivity to predict a subgroup of persistent wheezers. Nevertheless, continued exploration of machine learning methods, and the identification and integration of novel genomic biomarkers, may offer the potential to further improve the prediction of childhood asthma.

Chapter 6 Exploration of Childhood Asthma Genomic Biomarkers

6.1 Introduction

The CAPE and CAPP models developed in Chapter 5 were primarily developed using demographic, clinical and environmental predictors of childhood asthma reported in the literature, with some predictors shown to offer a large effect on future asthma development (discussed in Chapter 1 and Chapter 5). Furthermore, to maximise the potential clinical utility of the CAPE and CAPP models, only easily collectable and readily available data from health records or patient questionnaires was considered[261]. However, it is well-established that the consideration of biomarkers for asthma and allergy can facilitate the diagnosis, prediction and identification of asthma phenotypes, as well as aiding drug discovery and directing personalised asthma treatment[37,262,263]. For example, individuals with an early onset allergic asthma phenotype are more likely to present with elevated levels of blood eosinophils, serum IgE and FeNO[37]. Such biomarkers can be used as predictive markers to identify future asthmatics (with indications of specific asthma endotypes) who would likely benefit from therapeutics directly targeting the biological mechanisms that underpin these biomarkers[37]. For example, omalizumab, the first biologic to be approved for the treatment of severe asthma, is a monoclonal antibody which binds to circulating IgE; guidelines suggest the drug should be administered specifically to severe allergic asthmatics with elevated serum IgE levels and evidenced allergic sensitisation (e.g. through SPTs), and be avoided in those presenting with non-allergic asthma phenotypes[263].

A number of existing childhood asthma prediction models have incorporated biomarkers of asthma or allergy in the form of SPTs to measure allergic sensitisation, blood tests which measure blood eosinophilia; RAST tests (or other assays) to measure specific IgE levels; or FeNO to provide an indication of airway inflammation (Table 3.2). One study even assessed whether the addition of VOCs from exhaled breath condensates and gene expression data could improve predictions made by the original API[222]. In line with this, biological data in the form of SPTs (the only biomarker data collected from individuals up to the 4-year follow-up in the IOWBC which was also available in MAAS) was also considered during the development of the CAPE and CAPP models in Chapter 5. Indeed, the inclusion of SPTs offered a clear predictive benefit, with the exclusion of predictors derived from SPTs resulting in a 10% reduction in the discriminative performance (AUC) of the CAPP model.

Over the last few decades, advances in omic technologies have furthered the investigation of molecular and genomic biomarkers of asthma[264]. It is well-established that alongside environmental factors, there is a large genetic contribution towards the susceptibility of developing asthma[43,265,266]. As discussed in Chapter 1.1.6, genetic studies (from twin and candidate gene studies to recent well-powered genome-wide association studies using large sample sizes) have identified a multitude of genetic variants significantly associated with asthma, including variants which only confer modest effects[50,63]. Similarly, epigenetic studies, which investigate chemical changes to the DNA sequence (methylation or histone modifications) that may affect gene expression and potentially explain interactions between genetic and environmental factors, have identified differential methylation profiles between asthmatic and non-asthmatic individuals[267-269] (discussed in Chapter 1.1.6). The identification of genomic markers able to differentiate between asthmatic and non-asthmatic individuals from large-scale genome-wide and epigenome-wide association studies (i.e. GWAS and EWAS to identify disease associated SNPs and differentially methylated CpG sites, respectively) have the potential to be combined into genomic risk scores and exploited to further evaluate the risk of developing childhood asthma[53,63].

6.1.1 Objectives

In this chapter, genomic risk scores were developed to predict school-age asthma. Specifically, polygenic and epigenetic risk scores were constructed in the IOWBC using available genome-wide genotype and methylation data based on previously published lists of SNPs and CpGs found to be significantly associated with childhood asthma, respectively.

In line with the final two aims of this thesis (as described in Chapter 1.4), the performance of each genomic risk score to predict school-age asthma was assessed in the IOWBC and independently validated MAAS (where data was available) (Aim 6). These genomic risk scores were then integrated into the clinical models developed in Chapter 5 to assess whether the addition of asthma genomic biomarkers could offer any improvement in predictive performance (Aim 7).

6.2 Methods

6.2.1 Construction of a childhood asthma polygenic risk score

Of the 924 individuals with genotype data available in the IOWBC (data collection and quality control has been described in Chapter 2.1.1.2), 908 individuals had a defined asthma status at age 10 (141 asthmatic, 767 non-asthmatic) and were used to construct the polygenic risk score (PRS).

6.2.1.1 Candidate predictors for the PRS

To construct the PRS, 128 independent SNPs associated with asthma (annotated to 161 asthma target genes and 47 gene enriched pathways) were considered. These SNPs were identified from a recent study conducted by El-Husseini *et al*. which provided an updated summary of independent SNPs associated with asthma from published GWAS between 2007 and 2019[270]. In brief, SNPs with genome-wide significance ($p<5x10^{-8}$) were identified and tested for independence in European populations using the LDmatrix tool on LDlink[271]. The study considered SNPs to be independent if the linkage disequilibrium (LD) correlation (r^2) was less than 0.05.

The list of 128 SNPs was summarised from a combination of different asthma GWASs. Therefore, to construct the PRS specifically for childhood asthma, summary statistics for the 128 SNPs were extracted from a single GWAS study recently conducted by Ferreira *et al*.[272]. This was the largest GWAS that used data from UK Biobank (similar population to the IOWBC and MAAS) to identify SNPs associated with the most relevant childhood onset asthma phenotype. SNPs were included in the construction of the PRS if genotype data (post quality control) was available in the IOWBC and summary statistics were available in Ferreira *et al.*'s GWAS. Where data for a SNP was unavailable in either the IOWBC or within the GWAS summary statistics, the closest proxy SNP in high LD (r^2>0.8) within the European British in England and Scotland (GBR) population, with data available in the IOWBC, was sourced using the LDproxy tool on LDlink[273].

The summary statistic data extracted for each SNP from the GWAS included: i) effect size - to weight each SNP in the PRS, and ii) p-value - to determine the inclusion of SNPs in the score using the thresholding method. Where proxy SNPs were used, the effect size and p-value of the original SNP were used.

6.2.1.2 Calculation of the PRS

The childhood asthma PRS was constructed using the clumping and thresholding method using PRSice[274] and confirmed using the allele scoring '--score' command in PLINK (version 1.90)[208,209]. For this method, clumping is first performed to ensure that the SNPs which will be included in the PRS are independent by grouping nearby SNPs together and removing those in high LD. Next, thresholding is the process in which SNPs are included into the score based on whether they are below a pre-specified p-value threshold. Using the thresholding method, a number of scores are constructed across a range of p-value thresholds and the best score is selected based on a specified criteria. For both tools, the PRS was calculated as the sum of an individual's risk alleles weighted by the allele effect size for each SNP as estimated from the GWAS study (Equation 6.1).

$$PRS = \sum_{i}^{N} X_i \beta_i$$

Equation 6.1 Formula for calculating a weighted polygenic risk score

The PRS is calculated as the sum of all SNPs included in the score (N). For each SNP in the PRS (*i*), the dosage of the risk allele (*X*) is weighted by its GWAS effect size estimate (*β*).

By default, PRSice performs clumping and removes SNPs in high LD (r^2>0.1) within a 250kb window and calculates scores across all possible p-value thresholds. In contrast, when using PLINK, the parameters for clumping and thresholding need to be specified. As all SNPs considered for the PRS were already deemed independent, in low LD (r^2<0.05) with each other, it was not necessary to perform clumping as an additional step in PLINK. Furthermore, guided by a tutorial on calculating PRSs using PLINK and PRSice, a range of p-value thresholds (p-value<0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 1.0) were evaluated when using PLINK[274].

To select the best PRS from all the scores calculated from the thresholding method, PRSice uses Nagelkerke's R^2 goodness of fit statistic which evaluates how well the score explains the variance in the binary phenotype. Whilst Nagelkerke's R^2 was considered, in line with the selection of the best CAPE and CAPP models in Chapter 5, the final PRS was selected as the score which offered the highest AUC across 2000 bootstrapped samples.

6.2.2 Construction of childhood asthma epigenetic risk scores

Of the 765 individuals in the IOWBC with methylation data collected from Guthrie cards (data collection and quality control has been described in Chapter 2.1.1.3), 747 individuals also had a defined asthma status at age 10 (124 asthmatic, 623 non-asthmatic) and were used to construct the methylation risk score (MRS).

6.2.2.1 Candidate predictor for the MRS

To construct the MRS in the IOWBC, CpGs significantly associated with childhood asthma were extracted from a recent EWAS meta-analysis for childhood asthma published by Reese *et al*.[267] In this study, two separate EWAS meta-analyses for childhood asthma (7-17 years of age) were performed. The first was a prospective EWAS which used DNA methylation data from cord blood samples (newborn EWAS) whilst the second was a cross-sectional EWAS which used peripheral blood samples collected between 7-17 years (childhood EWAS). Two MRSs – a newborn MRS (nMRS) and childhood MRS (cMRS) - were constructed using significant CpGs identified from each EWAS meta-analysis. Of the 9 CpGs found to be significantly associated with asthma from the newborn EWAS, data for only 6 CpGs was available in the IOWBC after pre-processing of the methylation data. For the childhood MRS, the EWAS identified 164 CpGs associated with asthma, of which 157 has data available in the IOWBC.

To ensure that only independent CpGs were included in each MRS, the correlation between the CpGs considered for each model were evaluated. Due to the skewed distribution at some CpGs, Spearman's rank correlation coefficient was used to evaluate correlation between CpG sites (CpGs with $R^2>0.8$ were considered highly correlated). Independence between CpGs was also evaluated based on the distance between CpGs and their regional positions if found within the same CpG island; studies have identified that nearby CpGs (<2000 base pairs from each other) are often co-methylated and CpGs found within the same region of a CpG island are suggested to be non-independent[275,276]. Where correlated pairs of CpGs were identified, the CpG with the higher p-value reported from the EWAS meta-analysis was discarded.

A feature selection of the independent CpGs considered for each MRS was then performed by RFE using a random forest algorithm within a 5-fold cross validation. For feature selection, the beta values for each CpG were first standardised. In line with the feature selection performed in Chapter 5, the optimal subset of CpGs to include in each MRS was selected based on the average balanced accuracy score within a stratified five-fold cross-validation.

6.2.2.2 Calculation of the MRS

As no gold-standard method has yet been established for calculating MRSs[277], five different calculations identified from the literature were compared (Equation 6.2-6.6). For each MRS, the best score was selected as the calculation which offered the highest AUC across 2000 bootstrapped samples.

$$MRS\ 1 = \sum_{i}^{N} \beta_i w_i$$

Equation 6.2 MRS 1

Algorithm as reported by Fernandez-Sanles *et al.*[278]. Score is calculated as the sum of the beta value (*β*) multiplied by the effect size as reported in the EWAS meta-analysis (*w*), for all CpGs included in the score (*N*).

$$MRS\ 2 = \#X\ \{X\ \in M\}$$

Equation 6.3 MRS 2

Algorithm as reported by Guan *et al.*[279]. Score is calculated as the count (*#*) of the number of CpGs (*X*) that were hypermethylated or hypomethylated (*M*). A CpG was considered hyper (hypo)-methylated if methylation levels were in the upper (lower) quartile of the distribution among controls.

$$MRS\ 3 = \tfrac{1}{N}\textstyle\sum_{i}^{N} w_i(\tfrac{\beta_i - \mu_c}{\sigma_c})$$

Equation 6.4 MRS 3

Algorithm as reported by Yu *et al.*[280], where *N* denotes the number of CpGs considered in the MRS, *β* is the methylation (beta) value of the CpG and μ_c and σ_c are the mean methylation (beta) value and standard deviation among non-asthmatic controls, respectively. Each CpG in the score was weighted (*w*), with hyper(hypo)methylated CpGs assigned a weight of +1(-1).

$$MRS\ 4 = \frac{1}{N}\sum_{i}^{N} w_i \left(\frac{\beta_i - \mu_c}{\sigma_c}\right)$$

Equation 6.5 MRS 4

Algorithm as reported by Yu *et al.*[280], where *N* denotes the number of CpGs considered in the MRS, *β* is the methylation (beta) value of the CpG and μ_c and σ_c are the mean methylation (beta) value and standard deviation among non-asthmatic controls, respectively. Each CpG in the score was weighted (*w*) using the effect size reported in the EWAS meta-analysis.

$$MRS\,5 = \sum_{i}^{N} \frac{\beta_i}{\beta_N}(Y)$$

Equation 6.6 MRS 5

Algorithm as reported by Elliot *et al.*[281], where β_i is the methylation (beta) value of each CpG, β_N is the average effect size across all CpGs and *Y* denotes the difference between the CpG methylation value and the median methylation level reported among non-asthmatic controls (reference methylation beta value). For CpGs associated with an increased methylation level in asthmatics, *Y*= beta value – reference methylation beta value. For CpGs associated with a decreased methylation level in asthmatics, *Y*= reference methylation beta value –beta value.

6.2.3 Genomic biomarkers for childhood asthma prediction

To evaluate the predictive performance of each genomic biomarker, univariable prediction models were first developed. Next, the genomic biomarkers were integrated with the CAPE and CAPP models developed in Chapter 5 to assess whether the addition of each genomic marker could further improve the prediction of childhood asthma.

6.2.3.1 Univariable genomic prediction models

Prediction models using machine learning approaches were developed for each genomic risk score – a PRS prediction model, a nMRS prediction model and a cMRS prediction model. The same method used for the construction of the CAPE and CAPP models in Chapter 5 was applied. Individuals with data for each risk score and the asthma outcome were considered for the development of each model. The dataset was split into a training and holdout validation set (2:1 ratio, preserving class proportions). As each model consisted of a single feature (the genomic risk score), no feature selection was performed. Models were developed using all 7 machine learning algorithms and trained on the complete training dataset as well as the optimised training datasets which applied oversampling and/or undersampling to address the class imbalance (strategies i-iii, as detailed in Chapter 5.2.5). The model offering the best AUC in the hold-out validation set was considered as the best model. Performance measures were evaluated as described in Chapter 5.2.6.

6.2.3.2 Integration of the genomic biomarkers with the CAPE and CAPP models

The genomic biomarkers were integrated with the clinical models (CAPE/ CAPP) in a stepwise manner, whereby the following models were developed: i) clinical model plus PRS; ii) clinical model plus nMRS; iii) clinical model plus cMRS; iv) clinical model plus PRS and nMRS; and v) clinical model plus PRS and cMRS. The models were integrated by adding the relevant genomic risk scores as additional predictors to each clinical model's existing feature set.

The integrated CAPE and CAPP models were then retrained using the same algorithm and training dataset characteristics as identified in Chapter 5.3.5 and Chapter 5.3.6, respectively. Specifically, for each integrated CAPE model, the dataset of individuals with complete data for all features was split into a training and hold-out validation set (2:1 ratio, preserving class proportions) and the training dataset was undersampled to balance class proportions. In contrast, for each integrated CAPP model, the dataset of individuals with complete data for all features was split into a training and hold-out validation set (2:1 ratio, preserving class proportions) and the number of cases in the training dataset was oversampled by 300% and the number of controls further undersampled to balance class proportions. Both sets of integrated models were developed using support vector machine algorithms, with the hyperparameters for each model being tuned using a grid search (RBF kernel for the integrated CAPE model and linear kernel for the integrated CAPP models).

6.2.3.3 External validation of the genomic and integrated childhood asthma prediction models

The generalisability of the genomic risk scores and the integrated CAPE and CAPP models was assessed in the unselected MAAS cohort. Only individuals with complete data for the predictors in each model and the asthma outcome were used in the external validation analyses.

A PRS was calculated for each individual in MAAS using the SNPs included in the best PRS calculated in the IOWBC (described in Section 6.2.1.2). Even if proxy SNPs were used in the IOWBC, the presence of the original SNP detailed in the curated list of 128 independent asthma SNPs was first evaluated. Where SNPs were unavailable in MAAS, proxy SNPs were sourced as detailed in Section 6.2.1.1. If proxy SNPs were unavailable, SNPs with missing data were excluded from the cohort's PRS.

The MRSs and their subsequent integrated CAPE and CAPP models were unable to be validated in MAAS due to the unavailability of suitable methylation data samples (DNA methylation in MAAS was measured in cord blood samples using the Illumina 27K microarray).

6.3 Results

6.3.1 Childhood asthma polygenic risk score

From the list of 128 independent SNPs considered for inclusion in the PRS, GWAS summary statistics were unavailable for 3 SNPs. Of the remaining 125 SNPs, data for 105 SNPs were available in both the IOWBC and GWAS summary statistics and an additional 11 SNPs were accounted for using proxy SNPs, resulting in a total of 116 SNPs available to construct the PRS. In total, nine of the 128 SNPs were excluded due to the lack of proxy SNPs in high LD (R^2>0.8) being available in the IOWBC genotype data. Based on a ranking of the 125 SNPs by GWAS effect size, one of the nine excluded SNPs had the largest reported effect size (rs115468973 in HLA-DRB6) – the remaining 8 were middle-to-low ranking SNPs (further detail available at: https://doi.org/10.5258/SOTON/D1943).

Using PRSice, multiple scores were evaluated across a range of p-value thresholds. The best performing PRS consisted of 105 SNPs, calculated using a p-value threshold of p<0.047 (R^2=0.027, AUC=0.61) (Table A13, Figure 6.1A-B). Based on this 105-SNP PRS, individuals with asthma had a slightly higher mean PRS compared to those without asthma (Figure 6.1C). A quartile plot of all 908 individuals with a PRS calculated in the IOWBC demonstrated that an increasing PRS was associated with an increased risk of developing school-age asthma, with individuals in the highest quartile being 2.22 times more likely to develop asthma at age 10 compared to those in the lowest quartile (Figure 6.1D).

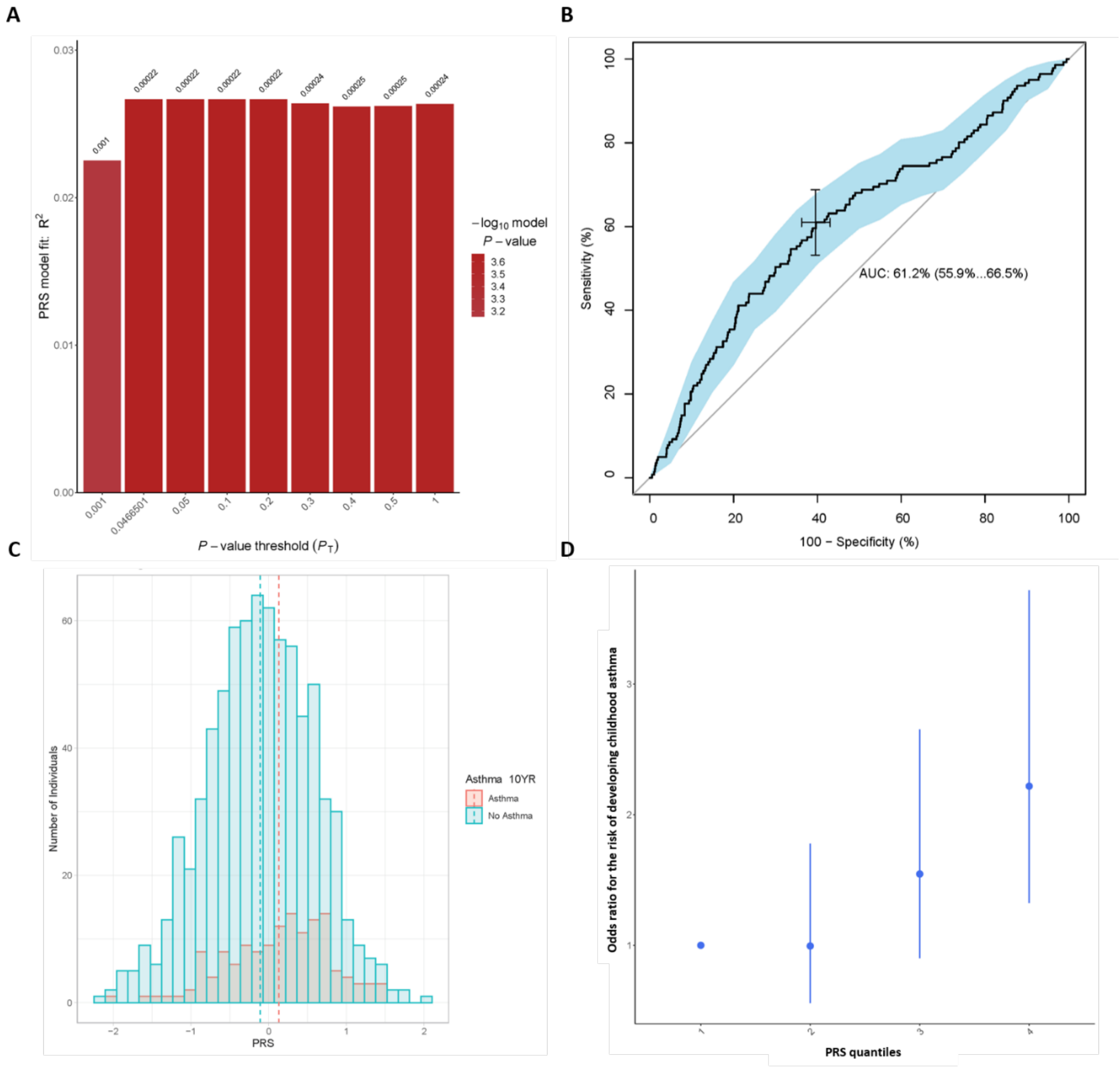


Figure 6.1 Evaluation of the best performing childhood asthma PRS in the IOWBC

Evaluation of PRSs constructed across a range of p-value thresholds identified a 105-SNP PRS, including SNPs with p-value less than 0.047, to explain the greatest degree of variance in the asthma phenotype (A). Discriminative ability of the 105-SNP PRS is presented using a ROC curve (B). The histogram presents the distribution of the PRS among asthmatic and non-asthmatic individuals in the IOWBC (C). The quantile plot illustrates the risk of developing school-age asthma at age 10 among individuals with increasing PRS, with respect to individuals with a PRS in the lowest quartile (D).

6.3.2 Childhood asthma methylation risk score

6.3.2.1 Newborn MRS

All six candidate CpGs considered for inclusion in the nMRS were deemed to be independent from one another - none of the CpGs were highly correlated (Figure A5), within 2000 base pairs from each other or found within the same CpG island. Feature selection using RFE selected all six CpGs for inclusion in the nMRS, offering the maximal balanced accuracy score of 0.58 (Figure 6.2A). Whilst all five calculations described in Table 6.1 offered similar discriminative performance, the best nMRS was calculated using MRS 1, achieving an AUC of 0.55 (95% CI: 0.50-0.60).

6.3.2.2 Childhood MRS

Of the 157 candidate CpGs considered for inclusion in the cMRS, 22 CpGs were within 2000 base pairs of another CpG; of which 12 CpGs were located within the same region of the same CpG island as another CpG. However, only one pair of CpGs were found to be highly correlated (r=0.93, Figure A6-7). Removal of the CpG with the higher p-value within this correlated pair of CpGs resulted in 156 independent CpGs for consideration in the feature selection. RFE selected 110 CpGs for inclusion in the cMRS, offering a balanced accuracy score of 0.58 (Figure 6.2B). Although all five calculations offered similar discriminative performance (Table 6.1), the best cMRS was calculated using MRS 2, achieving an AUC of 0.54 (95% CI: 0.49-0.59).

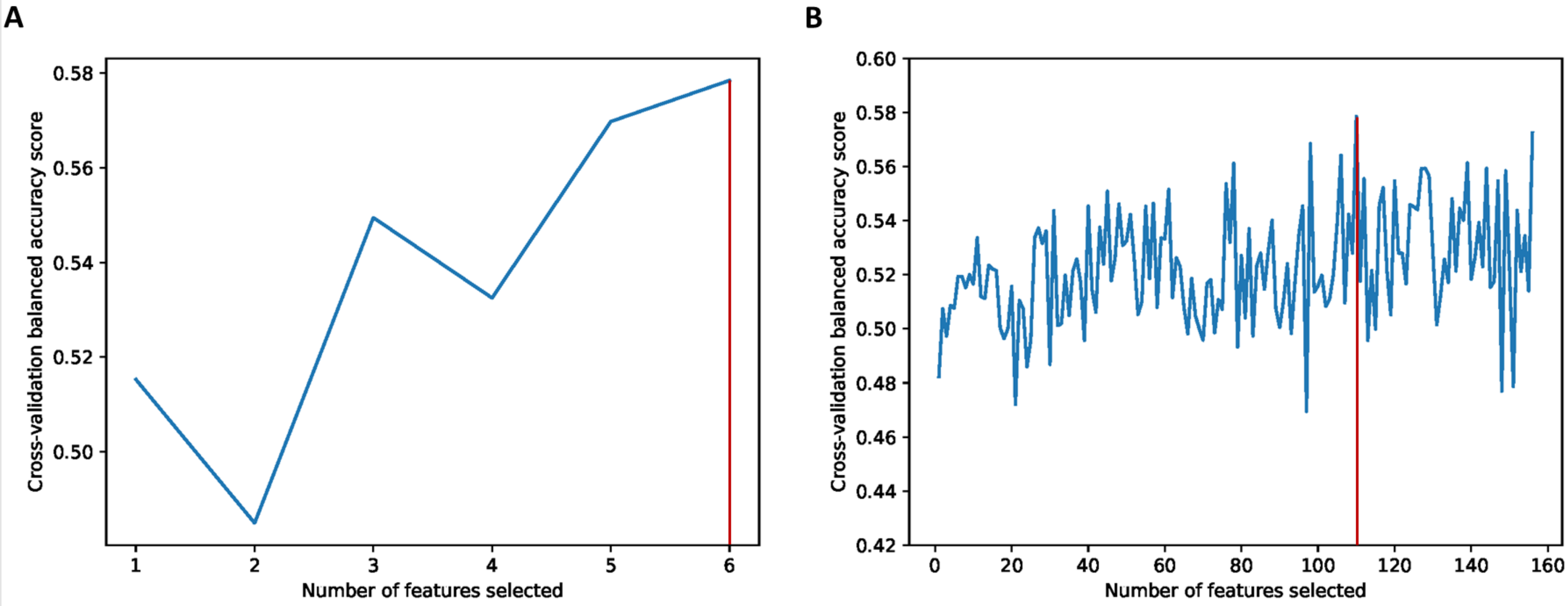


Figure 6.2 Feature selection for the identification of CpGs for inclusion in the newborn and childhood methylation risk scores

The optimal subset of features for inclusion in each model was identified as the subset which offered the best balanced accuracy score (red line).

Table 6.1 Performance of the newborn and childhood methylation risk scores calculated using five different equations

MRS	**Newborn MRS – 6 CpGs** **AUC (95% CI)**	**Childhood MRS – 110 CpGs** **AUC (95% CI)**
1	**0.55 (0.50-0.60)**	0.53 (0.48-0.59)
2	0.54 (0.48, 0.59)	**0.54 (0.49, 0.59)**
3	0.49 (0.44, 0.55)	0.53 (0.48, 0.59)
4	0.53 (0.48, 0.59)	0.53 (0.48, 0.59)
5	0.52 (0.47, 0.58)	0.53 (0.48, 0.59)

The scores offering the best discriminative performance, and which were used in subsequent analyses, are highlighted in bold.

6.3.3 Genomic biomarkers for childhood asthma prediction

6.3.3.1 PRS machine learning model

Of the 908 individuals with a PRS, 641 individuals were allocated to the training dataset (16% asthmatic) and 267 allocated to the hold-out validation set (16% asthmatic). The best performing PRS model was developed using an SVM classifier (linear kernel, C=0.15), trained on the complete training dataset, oversampled by 100% and undersampled to balance class proportions (n=412, 206 asthmatics and 206 non-asthmatics). The model performed with an AUC=0.64 (Figure 6.3). Based on the threshold cut-off that maximised the Youden's Index (threshold=0.53), classifications of asthma were made and performance measures evaluated (Table 6.2). Based on this threshold, the PRS model demonstrated moderate performance, with good specificity (76%) but modest sensitivity (identifying only half of future asthmatics) in the holdout validation set.

To validate the PRS in MAAS, data was available for 102 of the 105 SNPs included in the PRS (10 SNPs were accounted for by proxy SNPs, further detail available at: https://doi.org/10.5258/SOTON/D1943). 807 and 767 individuals had PRS data and a defined asthma status at 8 and 11 years of age, respectively. The PRS model demonstrated good generalisability to predict asthma (AUC=0.61) at both ages 8 and 11 in MAAS, offering a similar AUC as reported in the IOWBC (Figure 6.3, Table 6.2).

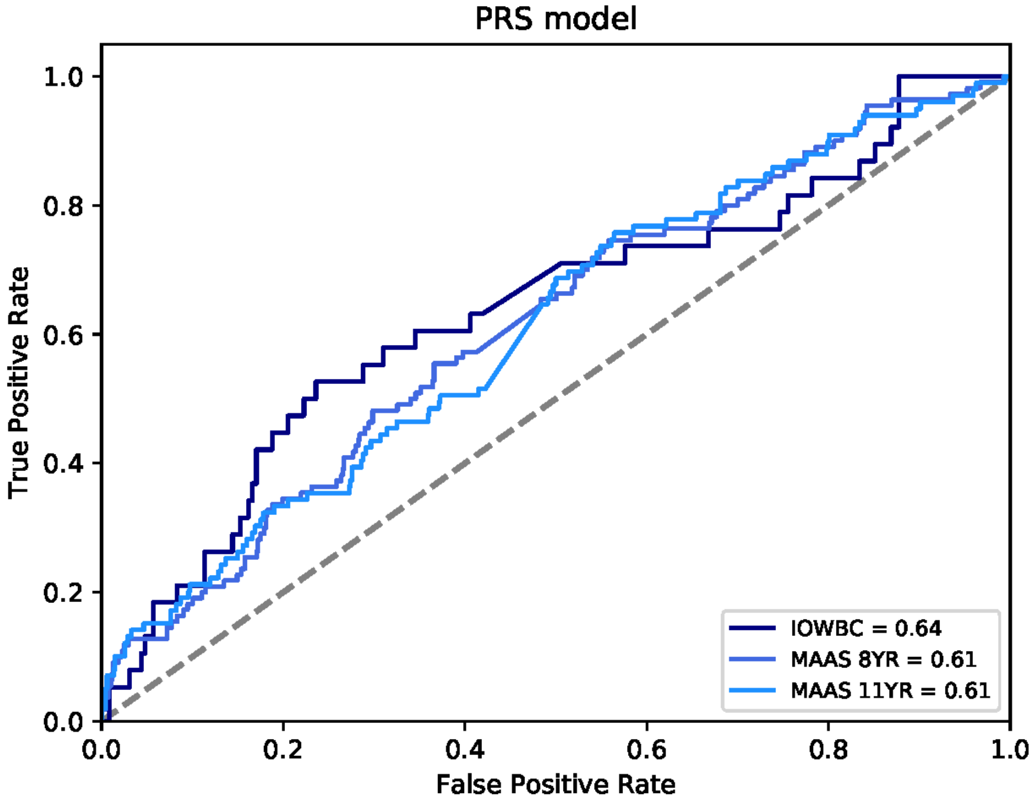


Figure 6.3 ROC curves comparing the performance of the univariate PRS machine learning model in the IOWBC and MAAS

Table 6.2 Performance of the PRS univariable machine learning childhood asthma prediction model in the IOWBC and MAAS

	Dataset	Sample size (# asthmatic)	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score
IOWBC: 10 years	Training	412 (206)	0.56	0.59	0.60	0.52	0.56	0.57	1.25	0.77	0.58
	Testing	267 (38)	0.65 (0.56-0.73)	0.64 (0.54-0.73)	0.53 (0.37-0.68)	0.76 (0.71-0.82)	0.27 (0.20-0.35)	0.91 (0.88-0.94)	2.23 (1.48-3.25)	0.62 (0.42-0.84)	0.36 (0.26-0.45)
MAAS: 8 years	Unselected	807 (110)	0.57 (0.53-0.62)	0.61 (0.56-0.67)	0.34 (0.25-0.43)	0.81 (0.78-0.84)	0.22 (0.16-0.27)	0.89 (0.87-0.90)	1.75 (1.24-2.33)	0.82 (0.71-0.94)	0.26 (0.20-0.33)
MAAS: 11 years	Unselected	767 (99)	0.57 (0.52-0.62)	0.61 (0.55-0.67)	0.33 (0.25-0.43)	0.80 (0.77-0.83)	0.20 (0.16-0.26)	0.89 (0.87-0.90)	1.67 (1.22-2.31)	0.83 (0.71-0.94)	0.25 (0.19-0.32)

The PRS univariable model was developed using an SVM classification algorithm using a linear kernel (C=0.15). The model was trained on the complete training dataset, with cases oversampled by 100% and controls under-sampled to obtain a 1:1 class ratio.

Performance measures in the IOWBC holdout validation set and unselected MAAS dataset are evaluated using a classification threshold of 0.53.

6.3.3.2 MRS machine learning models

Of the 747 individuals with MRSs calculated, 508 were allocated into the training dataset (18% asthmatic) and 239 allocated to the hold-out validation set (14% asthmatic). The best performing nMRS model was developed using a KNN algorithm (*k*=9), trained on the complete training set, oversampled by 100% and undersampled to balance class proportions (n=360, 180 asthmatics and 180 non-asthmatics). The model offered modest performance with an AUC=0.57 (Figure 6.4, Table 6.3). In contrast, a MLP classifier training on the complete training dataset, oversampled 150% and undersampled to balance class proportions offered the best performance for the cMRS model (AUC=0.63, Figure 6.4, Table 6.3).

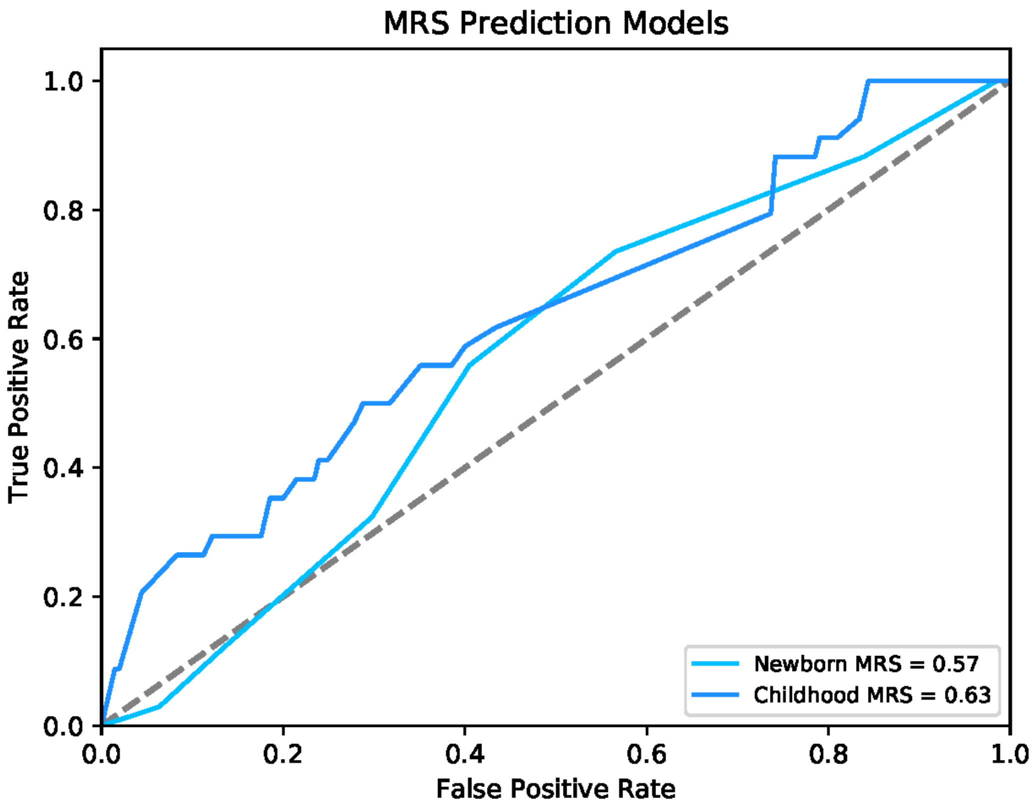


Figure 6.4 ROC curves comparing the performance of the univariable MRS machine learning models in the IOWBC

Table 6.3 Performance of the univariable newborn and childhood MRS machine learning childhood asthma prediction models in the IOWBC

	Dataset	Sample size (# asthmatic)	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score
Newborn MRS	Training‡	360 (180)	0.67	0.73	0.70	0.64	0.66	0.68	1.94	0.47	0.68
	Testing	239 (34)	0.58 (0.50-0.66)	0.57 (0.47-0.66)	0.74 (0.56-0.88)	0.43 (0.36-0.50)	0.18 (0.14-0.21)	0.91 (0.86-0.95)	1.30 (1.00-1.62)	0.61 (0.29-1.00)	0.29 (0.23-0.34)
Childhood MRS	Training‡	450 (225)	0.59	0.62	0.60	0.58	0.59	0.59	1.44	0.69	0.59
	Testing	239 (34)	0.61 (0.52-0.70)	0.63 (0.53-0.74)	0.50 (0.32-0.68)	0.71 (0.65-0.78)	0.22 (0.16-0.30)	0.90 (0.87-0.93)	1.74 (1.13-2.55)	0.70 (0.46-0.94)	0.31 (0.22-0.41)

The newborn MRS was developed using a KNN algorithm, trained on the complete training dataset, with cases oversampled by 100% and controls undersampled to balance class proportions. The hyperparameters of the model were: 'n_neighbours': 9, 'p': 1, 'weights': 'uniform'.

The childhood MRS was developed using a MLP algorithm, trained on the complete training dataset, with cases oversampled by 150% and controls undersampled to balance class proportions. The hyperparameters of the model were: 'activation': 'tanh', 'alpha': 1e-07, 'hidden_layer_sizes': (6, 6), 'learning_rate': 'constant', 'learning_rate_init': 0.1, 'solver': 'lbfgs'.

Performance was evaluated at the classification thresholds that maximised the Youden’s Index (newborn MRS=0.44 and childhood MRS=0.58).

6.3.3.3 Integration of genomic biomarkers to the CAPE and CAPP models

In the IOWBC, individuals with complete data for all predictors included in the integrated models were used. For the integration of the PRS with the CAPE model, applying the same characteristics to the training dataset (complete training dataset, undersampled to balance class proportions) resulted in a very small training dataset unsuitable for model training. Therefore, the number of asthma cases was oversampled by 100% prior to undersampling in order to obtain a dataset of similar size to the CAPE training dataset detailed in Chapter 5.3.6.

Overall, for both the CAPE and CAPP models, the integration of each genomic biomarker (either individually or in combination) did not significantly improve the discriminative performance of the models in the IOWBC (Table 6.4-6.5, Figure 6.5). However, marginal improvement in model discrimination was observed for the CAPE model upon the integration of the cMRS (AUC=0.75 vs 0.71). Similarly, for the CAPP model, marginal improvement was observed upon the integration of both the PRS and cMRS (AUC=0.84 vs 0.82).

Whilst the models integrated with the nMRS or cMRS were unable to be assessed in MAAS, the CAPE and CAPP models integrated with the PRS data were able to be replicated (Table 6.4-6.5). For the CAPE model, the addition of the PRS resulted in a slight decrease in model performance in the holdout validation set, and this was replicated when predicting asthma at age 8 in MAAS. When predicting asthma at age 11, the integrated CAPE model demonstrated equivalent performance to the original CAPE model (AUC=0.71) (Figure A8). However, for the CAPP model, whilst the addition of the PRS resulted in a slight reduction in AUC in the IOWBC holdout validation set, slight improvements in AUC were observed when predicting asthma at age 8 (AUC: CAPP=0.83 vs CAPP+PRS=0.85) and 11 years in MAAS (AUC: CAPP=0.79 vs CAPP+PRS=0.81) (Figure A9).

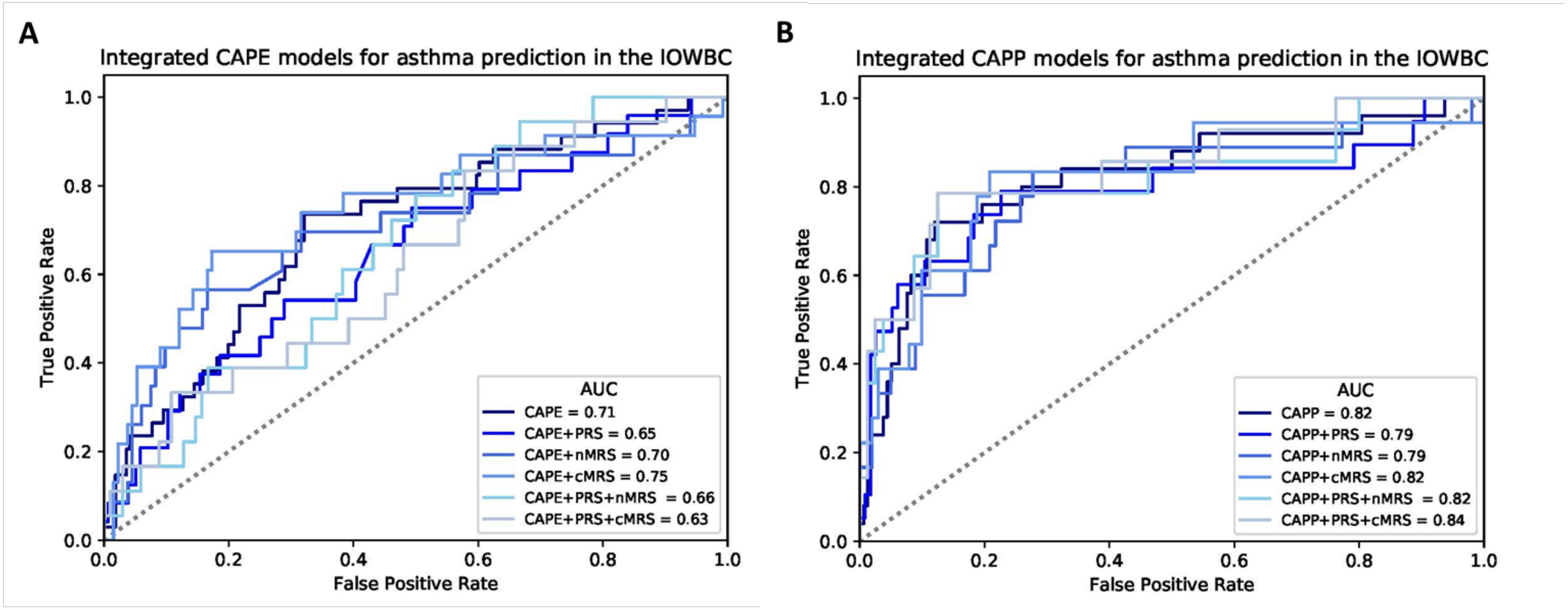


Figure 6.5 ROC curves comparing the performance of all integrated CAPE and CAPP models in the IOWBC

Table 6.4 Performance of all integrated CAPE models in the IOWBC and MAAS

	Dataset	Sample size (# asthmatic)	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score
CAPE	Training	136 (68)	0.65	0.76	0.56	0.75	0.69	0.63	2.24	0.59	0.62
	Testing	255 (34)	0.71 (0.62-0.78)	0.71 (0.61-0.80)	0.74 (0.56-0.88)	0.68 (0.62-0.74)	0.26 (0.21-0.32)	0.94 (0.91-0.97)	2.29 (1.69-3.01)	0.39 (0.18-0.63)	0.38 (0.31-0.46)
	MAAS 8YR	322 (38)	0.67 (0.60-0.74)	0.71 (0.63-0.79)	0.84 (0.71-0.95)	0.51 (0.45-0.56)	0.19 (0.16-0.21)	0.96 (0.93-0.99)	1.71 (1.40-2.03)	0.31 (0.10-0.57)	0.30 (0.26-0.35)
	MAAS 11YR	299 (32)	0.68 (0.60-0.74)	0.71 (0.62-0.79)	0.84 (0.72-0.97)	0.51 (0.45-0.57)	0.17 (0.14-0.20)	0.96 (0.94-0.99)	1.72 (1.39-2.05)	0.31 (0.07-0.58)	0.28 (0.24-0.33)
CAPE+PRS	Training	186 (49)	0.65	0.75	0.61	0.69	0.67	0.64	2.00	0.56	0.64
	Testing	180 (24)	0.63 (0.53-0.71)	0.65 (0.52-0.76)	0.75 (0.58-0.92)	0.51 (0.43-0.58)	0.19 (0.15-0.23)	0.93 (0.88-0.97)	1.52 (1.11-1.93)	0.49 (0.17-0.88)	0.30 (0.23-0.36)
	MAAS 8YR	270 (33)	0.54 (0.45-0.62)	0.65 (0.53-0.77)	0.70 (0.55-0.85)	0.38 (0.31-0.44)	0.13 (0.11-0.16)	0.90 (0.85-0.95)	1.12 (0.84-1.40)	0.81 (0.40-1.29)	0.23 (0.18-0.27)
	MAAS 11YR	266 (29)	0.61 (0.53-0.68)	0.71 (0.61-0.80)	0.83 (0.69-0.97)	0.39 (0.33-0.45)	0.14 (0.12-0.17)	0.95 (0.91-0.99)	1.35 (1.10-1.62)	0.44 (0.10-0.83)	0.24 (0.20-0.28)
CAPE+ nMRS	Training	92 (46)	0.72	0.75	0.61	0.83	0.78	0.68	3.50	0.47	0.68
	Testing	156 (23)	0.70 (0.60-0.81)	0.70 (0.56-0.82)	0.57 (0.35-0.78)	0.83 (0.77-0.89)	0.37 (0.26-0.50)	0.92 (0.88-0.96)	3.41 (1.99-5.78)	0.52 (0.26-0.77)	0.45 (0.31-0.59)

	Dataset	Sample size (# asthmatic)	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score
CAPE+ cMRS	Training	92 (46)	0.71	0.76	0.59	0.83	0.77	0.67	3.38	0.50	0.67
	Testing	156 (23)	0.74 (0.63-0.83)	0.75 (0.62-0.86)	0.65 (0.43-0.83)	0.83 (0.76-0.89)	0.39 (0.29-0.52)	0.93 (0.90-0.97)	3.77 (2.31-6.20)	0.42 (0.21-0.67)	0.49 (0.35-0.62)
CAPE+ PRS+ nMRS	Training	140 (70)	1.00	1.00	1.00	1.00	1.00	1.00	-	0.00	1.00
	Testing	120 (18)	0.64 (0.53-0.74)	0.66 (0.54-0.78)	0.78 (0.61-0.94)	0.50 (0.40-0.60)	0.22 (0.16-0.27)	0.93 (0.87-0.98)	1.56 (1.11-2.13)	0.44 (0.10-0.86)	0.34 (0.26-0.42)
CAPE+ PRS+ cMRS	Training	140 (70)	1.00	1.00	1.00	1.00	1.00	1.00	-	0.00	1.00
	Testing	120 (18)	0.63 (0.52-0.72)	0.63 (0.49-0.76)	0.83 (0.67-1.00)	0.42 (0.32-0.52)	0.20 (0.16-0.25)	0.93 (0.86-1.00)	1.44 (1.07-1.85)	0.40 (0.00-0.89)	0.33 (0.26-0.39)

Performance of each model was evaluated at the classification thresholds that maximised the Youden’s Index (CAPE+PRS=0.476, CAPE+nMRS=0.52, CAPE+cMRS=0.53, CAPE+PRS+nMRS=0.16, CAPE+PRS+cMRS=0.19).

- Unable to be calculated

Table 6.5 Performance of all integrated CAPP models in the IOWBC and MAAS

	Dataset	**Sample size (# asthmatic)**	**Balanced Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**LR+**	**LR-**	**F_1 Score**
CAPP	Training‡	408 (204)	0.78	0.85	0.80	0.77	0.78	0.79	3.47	0.26	0.79
	Testing	183 (25)	0.80 (0.70-0.89)	0.82 (0.71-0.91)	0.72 (0.52-0.88)	0.88 (0.83-0.92)	0.47 (0.38-0.62)	0.95 (0.92-0.98)	5.99 (3.79-10.11)	0.32 (0.13-0.54)	0.56 (0.45-0.70)
	MAAS 8YR	282 (33)	0.73 (0.64-0.81)	0.83 (0.75-0.90)	0.55 (0.36-0.70)	0.91 (0.88-0.95)	0.45 (0.33-0.59)	0.94 (0.92-0.96)	6.17 (3.64-10.69)	0.50 (0.33-0.69)	0.49 (0.36-0.62)
	MAAS 11YR	267 (29)	0.73 (0.63-0.82)	0.79 (0.68-0.88)	0.55 (0.38-0.72)	0.90 (0.87-0.94)	0.41 (0.29-0.55)	0.94 (0.92-0.96)	5.71 (3.44-9.85)	0.50 (0.30-0.71)	0.47 (0.33-0.62)
CAPP+PRS	Training‡	304 (152)	0.81	0.90	0.82	0.80	0.81	0.82	4.17	0.22	0.81
	Testing	134 (19)	0.78 (0.68-0.87)	0.79 (0.63-0.91)	0.79 (0.58-0.95)	0.77 (0.70-0.85)	0.37 (0.28-0.47)	0.96 (0.92-0.99)	3.49 (2.34-5.45)	0.27 (0.07-0.53)	0.50 (0.39-0.62)
	MAAS 8YR	239 (29)	0.77 (0.74-0.83)	0.85 (0.77-0.92)	0.97 (0.90-1.00)	0.61 (0.55-0.68)	0.25 (0.22-0.29)	0.99 (0.98-1.00)	2.47 (2.09-3.00)	0.06 (0.00-0.18)	0.40 (0.36-0.45)
	MAAS 11YR	238 (27)	0.70 (0.62-0.78)	0.81 (0.72-0.89)	0.81 (0.67-0.96)	0.59 (0.53-0.66)	0.20 (0.17-0.25)	0.96 (0.93-0.99)	2.00 (1.54-2.57)	0.31 (0.06-0.59)	0.33 (0.26-0.39)
CAPP+ nMRS	Training‡	280 (140)	0.85	0.92	0.82	0.88	0.87	0.83	6.76	0.20	0.85
	Testing	119 (18)	0.78 (0.67-0.87)	0.79 (0.65-0.91)	0.83 (0.67-1.00)	0.72 (0.64-0.81)	0.35 (0.27-0.44)	0.96 (0.92-1.00)	3.01 (2.08-4.49)	0.23 (0.00-0.50)	0.49 (0.39-0.60)

	Dataset	Sample size (# asthmatic)	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score
CAPP+ cMRS	Training‡	280 (140)	0.78	0.87	0.66	0.89	0.86	0.73	6.20	0.38	0.75
	Testing	119 (18)	0.81 (0.71-0.90)	0.82 (0.69-0.93)	0.83 (0.67-1.00)	0.79 (0.71-0.87)	0.42 (0.33-0.54)	0.96 (0.93-1.00)	4.01 (2.70-6.47)	0.21 (0.00-0.46)	0.56 (0.44-0.67)
CAPP+ PRS+ nMRS	Training‡	216 (108)	0.81	0.89	0.86	0.76	0.78	0.85	3.58	0.18	0.82
	Testing	94 (14)	0.83 (0.71-0.93)	0.82 (0.66-0.95)	0.79 (0.57-1.00)	0.88 (0.80-0.94)	0.52 (0.38-0.71)	0.96 (0.92-1.00)	6.29 (3.52-13.71)	0.24 (0.00-0.50)	0.63 (0.47-0.79)
CAPP+ PRS+ cMRS	Training‡	216 (108)	0.81	0.89	0.77	0.85	0.84	0.79	5.19	0.27	0.80
	Testing	94 (14)	0.83 (0.71-0.93)	0.84 (0.70-0.95)	0.79 (0.57-1.00)	0.88 (0.80-0.94)	0.52 (0.38-0.70)	0.96 (0.92-1.00)	6.29 (3.52-13.33)	0.24 (0.00-0.50)	0.63 (0.47-0.79)

Performance of each model was evaluated at the classification thresholds that maximised the Youden's Index (CAPE+PRS=0.38, CAPE+nMRS=0.31, CAPE+cMRS=0.34, CAPE+PRS+nMRS=0.59, CAPE+PRS+cMRS=0.60).

6.4 Discussion

6.4.1 Summary of findings

To account for potential genetic and epigenetic contributors that may improve the prediction of childhood asthma, a polygenic risk score and two variations of methylation risk scores for childhood asthma were developed. Univariable prediction models developed using machine learning classification algorithms for each genomic biomarker offered limited predictive ability. Furthermore, the integration of genetic and epigenetic risk scores with the CAPE and CAPP models did not offer any substantial improvement in predictive power.

6.4.2 Comparison with existing studies

A number of studies have used genomic risk scores for the prediction of health outcomes[282-284], including asthma and other allergic diseases[285-288]. Specifically for childhood asthma, Spycher *et al*. used SNPs found to be associated with childhood asthma (before 16 years of age) from one of the largest childhood asthma GWASs to date (from the GABRIEL consortium) to predict a number of asthma and non-asthma related phenotypes[289]. Whilst the PRS developed for childhood asthma unsurprisingly demonstrated poor power to predict non-asthma related phenotypes (high systolic blood pressure, high IQ and high body height), the PRS also only demonstrated modest performance to predict various asthma and wheeze phenotypes (AUC<0.60). Similarly, Belsky *et al.* also used the GABRIEL Consortium childhood asthma GWAS to construct a 15-SNP PRS among individuals enrolled in the Dunedin Multidisciplinary Health and Development Study[53]. Individuals at high genetic risk based on this PRS (above the median PRS value) were more likely to develop early onset childhood asthma (before 13 years of age, hazard ratio=1.12 [1.01-1.26]) and be at greater risk of developing life-course persistent asthma (onset before age 13 with recurrence up to 38 years, risk ratio=1.36 [1.14-163]). Further biological profiling of these asthmatics revealed that those at high genetic risk were also more likely to be atopic, have airway hyper responsiveness and incomplete reversible airflow, miss school or work due to asthma, and be hospitalised due to breathing problems. Despite these trends, this PRS was only able to offer moderate discriminative ability to predict future asthma (AUC=0.61). These findings correspond with the moderate predictive performance of the childhood asthma PRS constructed in this thesis, and may be a result of similar SNPs being included in the PRSs - all three PRSs included SNPs which were annotated to known asthma genes, specifically, the 17q12-21 loci, *IL33* and *IL1RL1* genes.

The construction of epigenetic (specifically methylation) risk scores are a relatively novel area of research, with one of the most prominent examples of MRSs being epigenetic clocks, shown to outperform other biological predictors of age[290,291]. MRS have also been used to provide insight into environmental factors, such as predicting an individual's smoking status[283,292,293]. MRSs have further been applied to predict a number of disease outcomes and have demonstrated good predictive ability. For example, a 16-CpG MRS offered very good performance to predict colorectal cancer (AUC=0.82)[294]. Similarly, a MRS of 75,000 CpGs demonstrated good discriminative ability to predict the development of major depressive disorder six years later, and was further shown to outperform a 27-variable clinical, demographic and lifestyle model, a 500–SNP PRS and a 5-feature biomarker risk score[295]. In contrast to the good predictive power reported by these MRSs, the performance of the nMRS and cMRS developed in this thesis was poor. Whilst much of the limited performance of the MRSs likely stems from study limitations (discussed in Chapter 6.4.3.2), it is also possible that DNA methylation alone cannot capture the heterogeneity of asthma compared to other disease outcomes. As the MRSs described in this thesis are the first known MRSs to be developed for asthma, further evaluation of the latter explanation was not possible.

A few studies have explored the integration of different data types in an attempt to improve the prediction of disease outcomes[280,283,296]. For example, Hamilton *et al*. demonstrated that the combination of phenotypic BMI and a MRS for BMI could explain a greater proportion of variation across a number of disease biomarkers (e.g. HbA1c, triglyceride levels, high density lipoprotein (HDL) cholesterol and HDL ratio) compared to phenotypic BMI data alone[296]. Similarly, for the prediction of lung cancer, the predictive performance based on the number of packs smoked per year was significantly improved upon with the addition of either a PRS or MRS, with further improvements with the integration of both genomic biomarkers (improvement in AUC from 0.78 to 0.81, and net reclassification improvement of 14%)[280]. Interestingly, the incremental integration of these different predictors revealed that the addition of the MRS contributed to a greater improvement in performance compared to the PRS. This supports the idea that whilst an individual's genetic profile can be predictive of disease, gene-environment interactions (captured through changes in DNA methylation levels) can offer significant contributions towards an individual's disease outcome[63]. However, studies exploring similar joint predictive modelling for childhood asthma are limited. Only one known study has explored the joint contribution of a genomic risk score with a personal and environmental risk score for childhood asthma[61]. This study used data from the PIAMA cohort to construct a personal and environmental risk score and two PRSs. Whilst one PRS used data for 22 SNPs significantly associated with asthma from a multi-ancestry GWAS meta-analysis conducted by the Trans-National Asthma Genetic Consortium (TAGC); the other PRS used 133 SNPs associated with allergic disease (asthma, eczema or hay fever) from a GWAS meta-analysis conducted within the SHARE consortium. Similar to findings reported in Chapter 6.3.3.3, the addition of neither PRS was able to significantly improve upon the personal and environmental score alone (AUC: personal and environmental score=0.65; addition of TAGC PRS=0.66; addition of SHARE PRS=0.65). This lack of predictive improvement was also replicated upon validation in the BAMSE cohort. One explanation for the lack of predictive improvement offered by the PRS may be due to the genetic contribution to asthma already being accounted for by predictors included in the clinical models. Whilst this may be plausible for the risk score conducted in the PIAMA cohort which incorporated a predictor of family history, such a predictor was not included in the CAPE or CAPP models. Hence, further exploration into this is needed. However, it is possible that the lack of predictive improvement may stem from limitations of the PRS itself.

6.4.3 Selection of genomic markers

6.4.3.1 Genetic variants included in the PRS

Previously developed asthma PRSs were developed using single large-scale GWAS studies[53,61,289]. A gold standard approach for constructing a PRS with good predictive potential would be to use a list of SNPs associated with childhood asthma that has been replicated across a number of GWAS studies. However, this is not always feasible due to differences in GWAS studies with respect to sample size, study power, asthma definition and population characteristics. Even GWAS studies conducted in similar datasets can often be highly affected by small variations in asthma definitions or methodology. For example, in 2019 Ferreira *et al*.[272] and Pividori *et al*.[52] independently conducted similar large-scale GWASs using data from UK biobank to identify distinctions between early and adult onset asthma. Ferreira *et al*. identified 123 SNPs associated with childhood onset asthma, defined as a doctor diagnosis of asthma ever before age 19 (98 SNPs were replicated in a 23andme dataset). In contrast, Pividori *et al*. identified 61 SNPs associated with asthma, of which 23 were specific to childhood onset asthma (defined as a self-reported doctor diagnosis of asthma before age 12) and 37 were shared between childhood and adult onset asthma. Comparing the SNPs identified from these two studies, only 21 of the 98 replicated Ferreira *et al*. SNPs (or 24 of all 123 Ferreira *et al*. SNPs) were identified in the GWAS performed by Pividori *et al*. Selecting a single large-scale study from existing asthma GWASs to construct the PRS in the IOWBC was difficult due to inconsistencies in study characteristics such as: asthma definition, the age used to define childhood onset asthma and ethnicity – factors that would have undoubtedly affected the GWAS results. Therefore, in order to account for as many known genetic variants for asthma as possible, and not limiting findings to a single study, a curated list of independent SNPs (specific to European populations) summarised from the majority of asthma GWASs reported in the literature[270], with effect sizes extracted from a GWAS specifically for childhood onset asthma, was deemed most appropriate to use for the construction of the PRS in this thesis. This list of independent asthma SNPs has subsequently been annotated to 161 asthma target genes and 47 gene enriched pathways. The identified pathways include immune pathways well-established in the pathogenesis of asthma, driven by MHC-II, IL22, IL2, IL-4, IL-33 and IL-1RL1 signalling (a comprehensive evaluation has been reported by El-Husseini *et al*.[270]).

Nevertheless, it is important to acknowledge that the limited performance of the PRS may stem from weaknesses in SNP selection. Indeed, only 8 of the 116 SNPs considered for the development of the PRS were individually significantly associated with childhood asthma in the IOWBC; hence, it is probable that the large proportion of non-significant SNPs diluted the predictive potential of the significant SNPs. As the list of independent SNPs used to construct the PRS was established by summarising findings from GWASs that considered a variety of asthma phenotypes (e.g. asthma, childhood onset, adult onset, asthma and other allergic diseases), it is possible that SNPs less related to the development of asthma in childhood specifically were included in the PRS. Furthermore, given the strong genetic heritability for asthma uncovered from twin studies[265,266], it is possible that existing asthma GWASs do not fully account for the heritability of asthma. Indeed, many of the SNPs identified by existing GWASs are common variants, often present in both asthmatic and non-asthmatic individuals. The identification and inclusion of rare variants may capture more of the known heritability of asthma and improve the discriminative ability of future risk scores[64]. Additionally, further exploration into methods which may account for some of the missing heritability of asthma, such as the exploration into gene-gene interactions and accounting for the effects of shared environments through familial studies, is warranted[63,64].

6.4.3.2 CpGs included in the MRS

In contrast to the large number of GWASs conducted for asthma, a recent systematic review identified only 16 EWAS studies that have been performed for asthma[268]. Of the 12 studies specific to childhood asthma, only two were performed on sample sizes >150 individuals, suggesting that study power is a key limitation of many existing asthma EWASs. Furthermore, most studies evaluate DNA methylation levels using cord blood or childhood peripheral or whole blood samples, with a few small studies using tissue specifically relevant to the disease such as nasal or airway epithelial cells. Unlike genotype data which is largely unchanged throughout an individual's life, methylation levels are changeable, variable across different tissue types and can potentially be reversible in response to different environmental exposures[297]. As a result, the source of DNA methylation and the time point at which data was collected is extremely important. For example, none of the 9 differentially methylated CpGs identified from Reese *et al*.'s newborn EWAS meta-analysis (prospective EWAS using cord blood samples) were found to be significant among the 179 CpGs identified from the childhood EWAS (cross-sectional EWAS using peripheral blood samples at 7-17 years)[267].

In the IOWBC, DNA methylation data suitable for school-age asthma prediction purposes was only available from Guthrie cards (7 days after birth). However, none of the existing asthma EWASs were performed using Guthrie blood samples[268]. It was also not feasible to conduct a sufficiently powered EWAS within the IOWBC due to the low sample size of individuals with DNA methylation data in the IOWBC being inappropriate to establish a separate EWAS discovery dataset of sufficient size and power. Favouring EWASs using cord blood samples due to Guthrie blood samples being collected at a closer time point compared to childhood blood samples was deemed inappropriate; correlation analyses indicate that Guthrie blood samples are actually more similar to childhood/adult samples than cord blood samples[297]. Therefore, as the largest asthma EWAS to date, with the additional benefit of being a meta-analysis of 9 cohorts part of the PACE consortium, the Reese *et al*. study was used to construct the MRSs in this thesis. In an attempt to account for DNA methylation sample inconsistencies, rather than combining significantly associated CpGs between the two EWASs, two separate MRS using Guthrie blood samples were constructed based on findings from the newborn and childhood EWAS meta-analyses. Significant CpGs identified by Reese *et al*. have been annotated to asthma-related genes such as *ACOT7* and *IL5RA*, with genes enriching a number of immune pathways (a comprehensive analysis has been reported by Reese *et al*.[267]).

Unlike with the PRS, where a list of independent SNPs was already available, all significant CpGs identified from the EWASs were first evaluated for collinearity before undergoing feature selection to identify the optimal subset of CpGs for inclusion in the MRS (in line with a previous study by Deng *et al*.[298]). Based on Reese *et al.*'s functional analysis, all 6 CpGs considered for the newborn MRS were near a transcription factor binding site[267]. Of the 179 CpGs identified from the childhood EWAS, 113 CpGs were found in DNAse hypersensitivity sites, but only 17 and 34 CpGs were localised to CpG islands and promoters, respectively. Whilst the consideration of functional CpGs alone, annotated to regulatory regions or transcription sites, was not performed in this thesis, it is possible that this could have improved the discriminative performance of the MRSs.

6.4.4 Construction of the genomic risk scores

The PRS was developed using the clumping-and-thresholding method, a robust method for constructing polygenic risk scores[299]. As described in Chapter 6.2.1, this method allows for a number of scores to be calculated using SNPs associated with the phenotype across a range of significance levels before selecting the score at the best p-value threshold (i.e. the score that best explains the phenotype of interest). Whilst some studies have advocated that a greater number of significant SNPs, if not all genotyped SNPs, should be included into PRSs, it has been suggested that the further addition of SNPs (which also confer modest effects individually) does not improve the discriminative performance of a PRS[289,300,301]. Rather than selecting SNPs identified from individual SNP tests, the robust selection of associated SNPs derived and replicated in large GWASs and/or meta-analyses are most likely to improve the performance of PRSs. It is also important to note that other methods for constructing PRSs, such as LDpred and LASSO, have been proposed[299]. Some studies have also suggested that effect size should be leveraged over p-value when selecting SNPs for inclusion in a PRS[302].

As previously mentioned, the construction of MRSs is a relatively novel field, with no gold-standard method yet established[277]. Therefore, a number of methods reported in the literature were applied in this study (see Equation 6.2-6.6). Calculating the sum of significant CpGs, weighted by their methylation level (beta value) is a method similar to that used to construct the PRS. Indeed, this calculation gave rise to the best performing newborn MRS. Interestingly, a simpler scoring method that merely weights CpGs based on whether they are hyper/hypomethylation gave rise to the childhood MRS with the best discriminative performance. Whilst different calculations gave rise to the best newborn and childhood MRSs, it is important to acknowledge that all MRS calculations offered very similar performance. The application of other methods, such as LASSO, which have already been suggested for PRSs, may hold promise for future asthma MRSs[277].

6.4.5 Integration of genomic biomarkers

To integrate the genomic risk scores with the CAPE and CAPP models, the risk scores were added as additional predictors to the models and redeveloped using the same training characteristics as the original CAPE and CAPP models. Comparing the performance of the original CAPE and CAPP models and the models integrated with the PRS and/or MRSs, the genomic biomarkers did not appear to offer any predictive benefit. Whilst this is similar to previous reports that adding genetic information does not improve the prediction of childhood asthma[61] (previous studies have not integrated methylation data to asthma prediction models), it is important to acknowledge potential methodological decisions that may have limited model performance. First, the reuse of the same training dataset characteristics and classification algorithm that offered the best performance for the clinical models may not offer the optimal performance for the newly integrated models. Second, the genomic risk scores were integrated with the clinical models by simply adding the genomic risk score as an additional predictor in the feature set. A new feature selection including the original candidate predictors and the genomic biomarkers may have resulted in a different optimal feature subset being selected for the CAPE and CAPP models. Furthermore, the simple addition of the genomic predictors to the clinical models was initially pursued due to its simplicity and similarity to methods utilised in previous studies[61,280]. However, using a more complex method to combine the clinical and genomic models may have improved the performance of the integrated models. For example, a stacked generalisation method, combining the CAPE or CAPP machine learning models and the individual genomic biomarker machine learning models, could have been adopted to construct a meta-model. Using this method, it is suggested that the combined model should perform with equal, if not superior, predictive power compared to the best performing single model[303,304]. In theory, such a method would have prevented reductions in model performance that were observed upon the integration of the genomic risk scores with the clinical models. Finally, it is possible that the integration of both MRSs within the same model could have improved performance. However, in a real-world setting, this would require blood samples to be collected at two different time-points. With the methodological limitations already discussed and the limited performance of each MRS individually, integrating both MRSs with the CAPE and CAPP models (with/without the PRS) was deemed unnecessary and would have been unlikely to have substantially improved model performance.

6.4.6 Strengths and limitations

Whilst the construction and integration of the genomic risk scores for asthma did not lead to significant improvements in the predictive performance of the CAPE and CAPP models developed in Chapter 5, the study did have a number of strengths. First, a PRS to predict childhood asthma was constructed using a comprehensive list of SNPs summarised across most asthma GWASs to date, with performance being replicated in an independent population. Second, this is the first known study to construct and evaluate the predictive potential of methylation risk scores for childhood asthma using data from a large EWAS meta-analysis. Third, this is also the first known study that directly compared different calculations for constructing MRSs and identified that all evaluated calculations offered equivalent performance to predict childhood asthma. Fourth, the incremental integration of genomic markers with the existing childhood asthma prediction models enabled a thorough evaluation of each genomic model's predictive capability individually as well as across all different combinations of data aggregation. The generalisability of these integrated models (clinical and genetic data only) was also confirmed in the independent MAAS cohort, with models demonstrating similar performance to that displayed in the developmental cohort (IOWBC).

However, this study did have a number of limitations. As previously discussed, the EWAS used to select CpGs for inclusion in the MRS was inappropriate due to differences in DNA methylation sample collection. Furthermore, DNA methylation data available in MAAS was inappropriate for comparison with the methylation data used in the IOWBC; in MAAS, available DNA methylation profiles were deduced from cord blood samples using the Illumina 27K microarray (lower coverage than the EPIC array). Consequently, none of the MRSs or integrated models using the MRSs were able to be replicated. In addition, whilst the PRS was constructed using SNPs associated with asthma in European populations, the Reese *et al*. EWAS used to construct the MRSs was performed among individuals of mixed ancestry – it is possible that a EWAS focused on European ancestry may have offered better predictions of childhood asthma in the predominantly Caucasian IOWBC. Finally, the development of the integrated machine learning models, particularly the models that integrated all data types, were limited by sample size. It is possible that improvements in asthma prediction with the integration of genetic and methylation risk scores may have been observed using larger datasets for model training and validation.

6.4.7 Conclusion

Using data integration approaches, novel genomic risk scores (PRS and two MRSs) for childhood asthma were combined with the previously developed CAPE and CAPP models. Whilst machine learning models developed for each genomic risk score offered moderate predictive performance individually, the integration of genetic and epigenetic data in the form of these genomic risk scores were unable to significantly improve upon the performance of the CAPE and CAPP models (which were developed using clinical and environmental data alone). Based on these results, the clinical application of the genomic biomarkers for asthma prediction appears limited. Future research into developing PRSs and MRSs for childhood asthma using datasets of larger sample sizes, novel methodologies and more appropriate data sources for CpG selection are warranted. Consideration of other omic data, such as transcriptomics and metabolomics, may also improve the predictability of childhood asthma[282]. Furthermore, despite their current limited predictive capabilities, genomic biomarkers for predicting childhood asthma may be used to gain further research insights. For example, extending research beyond gene-environment interactions, and further exploring PRS-environment interactions, may uncover non-linear effects of environmental risk factors and subsequently identify patient subgroups that may benefit from precision medicine[63,305]. Consideration of such interaction effects as predictors in future models could potentially improve childhood asthma predictions.

Chapter 7 Discussion and Future Work

7.1 Summary of thesis findings

As discussed throughout this thesis, the ability to predict which children will develop asthma at school-age is difficult due to the highly heterogeneous pathophysiology, presentation and risk factors of childhood asthma, combined with the use of non-specific definitions of the disease, particularly in early life. In line with the aims of this thesis, a number of insights into childhood asthma prediction have been identified, with in-depth discussions on the strengths and limitations of each analysis detailed within the relevant chapters.

First, from the systematic review conducted in Chapter 3, an evaluation of existing prediction models for childhood asthma was performed. Key challenges that can hinder the development, validation and ultimate clinical utility of the childhood asthma prediction models were uncovered, leading to a detailed discussion on recommendations for future studies. One recommendation for future research was the exploration of novel methods, such as machine learning approaches. Although childhood asthma prediction models developed using more complex machine learning approaches have recently emerged, the CAPE and CAPP models developed in Chapter 5 are the first models developed using machine learning methods to assess model generalisability through an external validation in an independent population. These models were able to outperform their existing benchmark regression-based models, demonstrate good generalisability to identify true future asthmatics and offer excellent sensitivity to predict a subgroup of individuals presenting with a persistent wheeze phenotype. This supports future exploration of machine learning approaches to improve model performance. However, despite their potential to improve predictive performance, complex machine learning algorithms have infrequently been applied in medical research due to their reputation as uninterpretable “black-box” models[167,168,206]. The application of SHAP illustrated one potential way to address the issue of model interpretability among such “black-box” models, obtaining both global (overall prediction model) and local (individual predictions) explanations.

With the opportunity to further improve the performance of these prediction models through the incorporation of asthma biomarkers, genomic risk scores using genotype and methylation data were developed in Chapter 6. These genomic risk scores only indicated limited capability to predict childhood asthma, with the PRS only able to offer moderate discriminative performance and the first known methylation risk scores for childhood asthma offering modest predictive performance (although this was likely due to significant methodological limitations as discussed in Chapter 6.4.3.2). The incremental integration of these genomic biomarkers with the CAPE and CAPP models further highlighted the limited predictive capability of the genomic risk scores. However, whilst these genomic biomarkers were not fruitful in improving asthma predictions within this thesis, it is possible that these scores may possess widespread research application, with the potential to encourage future personalised asthma research.

7.2 Implications of thesis findings

The findings of this thesis could have a number of potential implications for different stakeholders.

7.2.1 For patients and parents

Whilst it is important to acknowledge that the CAPE and CAPP models developed in this thesis do not offer perfect predictive performance, they were able to offer superior performance and generalisability compared to existing prediction models for childhood asthma. Therefore, similar to the encouraged use of the PARS model[234], there is potential for the widespread use of the CAPE and CAPP models by parents/carers in the future. The application of these models could provide parents insight into their child's risk of developing childhood asthma and provide explanations of what features contributed to each individual's predicted probability of developing asthma at school-age. Although interpretation of these models by parents/carers alone may not result in clinical intervention, it is possible that parents, who identify their child to be at high risk of developing asthma at school-age, may be more cautious of potential risk factors, or even be motivated to alter their lifestyles in an attempt to mitigate their child's potential risk e.g. by reducing their child's exposure to tobacco smoke and other avoidable sources of environmental pollutants.

However, future use of these models would first require the development of a user-friendly online tool.

7.2.2 For physicians

Similar to the potential use of the current CAPE and CAPP models by parents, there is potential for their use by physicians to identify high-risk individuals for further clinical monitoring, or to target asthma management or preventative strategies. However, the limited clinical utility of similar existing prediction models for childhood asthma thus far cannot be ignored. A key reason for this is the limited assessment of generalisability among existing models, thus limiting them as exploratory studies. In those that have undergone validation, many models show modest generalisability with a decline in the ability to identify true future asthmatics. Hence, these models are often unable to improve upon predictions made based on a physician's clinical judgement. The CAPE and CAPP models developed using machine learning methods address this issue by their ability to offer improved performance over existing models (including the ability for the CAPP model to identify true future asthmatics), with these findings being replicated in another UK birth cohort. Yet, it is important to acknowledge that these models do not claim to be superior to a physician's clinical judgement. Rather, the clinical use of these models could act as unbiased tools which have the potential to support a physician's decision-making. Similar to previous applications of the API[224], there is also potential for the CAPE and CAPP models, as well as the childhood asthma PRS, to be used to support the identification of high-risk individuals for inclusion in clinical trials for therapeutics or prevention studies.

Furthermore, the ability to trust and understand how predictions were made by a model has been a significant hurdle that has specifically hindered the application of "black-box" machine learning algorithms in healthcare[167,168,206]. Through the application of SHAP, a potential method to overcome these concerns is presented, hopefully encouraging future exploration of machine learning to solve healthcare problems and reinforcing the supportive potential of these models for physicians.

7.2.3 For researchers

The CAPE and CAPP models developed using machine learning methods demonstrated improved performance and generalisability over their regression-based benchmark models. This promotes the need for researchers to continue to explore the potential application of machine learning approaches for predicting childhood asthma. Limitations of the models developed within this thesis also highlight potential areas for improvement.

Furthermore, the limited performance of the genomic risk scores to predict childhood asthma, both alone as well as upon integration with the CAPE and CAPP models, could suggest that genotype and methylation data are poor genomic biomarker for childhood asthma. Indeed, with existing PRSs for childhood asthma reported to offer similar limited predictive performance[53,61,289], researchers are guided towards one of three conclusions: i) exploration into different methodologies and the identification of more predictive SNPs (e.g. identifying rare variants, performing burden tests or conducting pathway analyses rather than single SNP tests) are needed to develop more predictive genetic risk scores; ii) the genetic risk of asthma has already been accounted for by predictors in the CAPE and CAPP models; or iii) further research is needed to uncover and predict the 'missing heritability' of childhood asthma[62-64 306]. Moreover, despite substantial study limitations which may have limited their predictive capability, the development of the first MRSs for childhood asthma in this thesis encourages further exploration into novel risk scores of different omic datatypes for childhood asthma prediction. The comparison of different formula for calculating MRSs also provides researchers with important insight into current methodologies for constructing MRSs, specifically due to the novelty of MRSs in medical research[277].

7.3 Future work

The need for further research into machine learning, genomic risk scores and methods of data integration for the prediction of childhood asthma is evident and has been previously discussed. Alongside previously discussed insights and an acknowledgement of limitations (including time restrictions) which may have impacted the work conducted within this thesis, a number of areas for future work have been suggested:

1. **Further evaluation of model generalisability**

Particularly if the CAPE and CAPP models are to be considered for widespread use by parents and/or physicians in the UK, the assessment of model generalisability within other UK birth cohorts would be beneficial. Specifically, the CAPE and CAPP models developed in this thesis were derived and validated within cohorts which predominantly consisted of individuals of Caucasian ancestry. For a robust assessment of model generalisability, replication of the CAPE and CAPP models in populations of different ethnic, environmental (rural/urban living or different countries or continents) or genetic backgrounds, is warranted.

2. **Consideration of different machine learning and data integration approaches**

Although a comprehensive list of supervised machine learning algorithms and methods to optimise model training were considered during the development of the CAPE and CAPP models, it is possible that other machine learning algorithms, such as gradient boosting algorithms, could generate more accurate prediction models[118,161]. Similarly, the exploration of different methods to integrate predictors of multiple datatypes may have improved the predictive performance of the models developed in Chapter 6. Such data integration methods could include: model-based approaches (e.g. constructing different models for each datatype and developing a final meta-model using stacked generalisation); concatenation-based approaches (e.g. performing a feature selection on a single matrix of all candidate predictors (of all datatypes); or a deep-learning approach (e.g. using each hidden layer in a neural network for predictors of a different datatype)[161,303,307,308].

3. Exploration and use of better quality datasets

Whilst the IOWBC used to develop models in this thesis is a good quality dataset of moderate size and rich information related to the development of asthma[170], limitations in sample size, class imbalance and missing data led to compromises in study design which may have resulted in bias during feature selection, data leakage between model training and validation, as well as suboptimal model training and performance. For future studies to avoid such limitations, datasets of larger sample sizes should be used. Furthermore, notable improvements in model performance following the application of oversampling and undersampling techniques reinforces the need for future studies to actively address the issue of class imbalance, a common problem in healthcare datasets. The application of class imbalance techniques in studies developing regression-based models is a particularly important area of future research. Without such studies, it will be difficult to determine whether the improved performance and generalisability of the CAPE and CAPP models developed in this thesis was in fact due to the use of complex machine learning algorithms as initially hypothesised, or an artefact of the dataset upon being treated for class imbalance.

Furthermore, as outlined in Chapter 3, future studies which aim to develop and validate childhood asthma prediction models should focus on the use of good quality datasets – datasets of large sample sizes, offering rich information in terms of data diversity and completeness, and which use standardised definitions for predictors and outcomes of interest would be ideal to support the robust development and independent validation of future models. An example of such a dataset is the STELAR consortium, a harmonisation of data from five UK birth cohorts (including the IOWBC and MAAS)[187]. The initial plan for this thesis was to undergo model development using data from the largest STELAR cohort, the Avon Longitudinal Study of Parents and Children (ALSPAC), followed by independent validation to assess model generalisability across the remaining four independent STELAR cohort populations. However, due to challenges with data access and time constraints for the completion of this thesis, it was not possible to fulfil this plan. Yet, based on the methodology described and optimised within this thesis, there is the potential to exploit the rich dataset of the STELAR consortium to develop new prediction models for childhood asthma as well as other allergic diseases.

4. Further exploration of genomic biomarkers predictive of childhood asthma

The exploration of genomic risk scores for childhood asthma is in its infancy and warrants further research. As previously mentioned, the consideration of novel methods and/or the identification of more predictive SNPs will be needed to improve the predictive performance of future PRSs for childhood asthma. In addition, due to sampling limitations faced in this thesis, it is likely that redevelopment of the MRSs using samples collected at more appropriate time points would improve upon the performance of the current childhood asthma MRSs. However, as methylation risk scores are relatively novel, with few applications reported in the literature, further exploration into methodologies for constructing MRSs is also warranted.

Future research to improve the predictability of childhood asthma may also include the development of risk scores using other omic datatypes which may be capable of discriminating between asthmatic and non-asthmatic individuals (e.g. transcriptomics, proteomics and metabolomics)[264]. In addition, future identification and integration of significant genetic-environment interactions or transgenerational epigenetic signals which may account for some of the missing heritability of asthma may be of benefit.

5. Exploration of a more appropriate prediction outcome – asthma definition, prediction metrics, asthma phenotypes

For any prediction model to be of clinical value, it is important that the purpose and clinical relevance of the model is determined, and that appropriate performance metrics and classification thresholds are subsequently used to evaluate the ability to predict the outcome of interest. For example, the ability for a model to rule in asthma may be preferred if the intended goal of a prediction model is to identify all potential future asthmatics (with little risk towards those offered false positive predictions). However, if a prediction model is used as part of the inclusion criteria for a clinical trial of a preventative intervention with a high risk of adverse effects, the ability to rule out non-asthmatics may be preferred in order to minimise unnecessary risks towards individuals who may not be asthmatic at school-age. Currently, the performance of existing childhood asthma prediction models is often reported at a cut-off based on the Youden's index, however the optimal threshold can vary between models, impairing the direct comparison between models. In line with this, clinical and research experts need to provide insight into the appropriate performance metrics to consider in order to facilitate meaningful evaluations of individual models and aid comparisons between models.

In addition, whilst the CAPE and CAPP models were developed to predict childhood asthma, the outcome of childhood asthma is difficult to define and identifies a highly heterogeneous group of individuals[249]. Whilst this highlights the need for researchers to collaborate with physicians in order to deduce a consensus definition of childhood asthma, there is a strong argument to consider asthma, not as a single disease, but as an umbrella term for a number of respiratory conditions[1]. Furthermore, it is possible that well-established risk factors of childhood asthma may only be associated with specific subtypes of asthma and/or among individuals with certain profiles of genetic risk. These possibilities encourage further exploration into the effect of such interactions and for future prediction models to be developed for specific patient subgroups[305]. Therefore, rather than developing an all-encompassing asthma prediction tool, research into predicting specific 'asthmas' among distinct patient subgroups using machine learning approaches may offer greater predictive insight and clinical utility of future childhood asthma prediction models.

Appendix A Supporting Material

Table A1 Definitions of the 54 candidate predictors used to develop the asthma prediction models

Candidate predictor	Definition
Family History	
Maternal smoking at birth	Maternal smoking status during pregnancy
Paternal smoking at birth	Paternal smoking status during pregnancy
Maternal asthma	Maternal asthma ever
Maternal eczema	Maternal eczema ever
Maternal hay fever	Maternal hay fever status
Paternal asthma	Paternal asthma status
Paternal eczema	Paternal eczema status
Paternal hay fever	Paternal hay fever status
Parity	Position of child in the family
Maternal socioeconomic status	Maternal socioeconomic status
Prenatal and postnatal	
Maternal age	Maternal age at pregnancy
Prematurity	Gestation age
Delivery	Mode of delivery
Total breastfeeding	Total breastfeeding duration
Exclusive breastfeeding	Exclusive breastfeeding duration
Solid food introduction	Age, in months, at which solid foods were introduced to the child's diet
Birthweight	Birth weight (kg)
Sex	Child's gender
Season of birth	Season at the time of the child’s birth: autumn (September-November), winter (December-February), spring (March-May), summer(June-August)
Dog	Household pet dog during pregnancy
Cat	Household pet cat during pregnancy
Furry pet	Household furry pet during pregnancy - dog, cat or other animal
Early life (combination of 1-year and 2-year follow-ups)	
SDS BMI	Child's BMI at age 1, standardised against the British 1990 growth reference.

Wheeze	Occurrence of wheezing before age 2
Wheeze without cold	Likely occurrence of wheezing in the absence of a cold before age 2
Cough	Occurrence of cough before age 2
Nasal symptoms	Occurrence of nasal symptoms before age 2
Chest infection	Occurrence of chest infections before age 2
Nocturnal symptoms	Occurrence of nocturnal asthma symptoms before age 2
Eczema	Eczema status by age 2
Hay fever	Hay fever status by age 2
Atopy	Atopy status (sensitisation to one or more allergens) by age 2
Monosensitisation	Sensitisation to one allergen by age 2
Polysensitisation	Sensitisation to two or more allergens by age 2
Parental smoking	Household parental smoking status by age 2
Dog	Household pet dog by age 2
Cat	Household pet cat by age 2
Furry pet	Household furry pet (dog, cat or other animal) by age 2
Early life residence on a farm	Main residence on a farm in the first year of life
Preschool age (4-year follow-up)	
SDS BMI	Child's BMI at age 4 (standardised against the British Growth Reference)
Wheeze	Occurrence of wheezing at age 4
Wheeze without cold	Likely occurrence of wheezing in the absence of a cold at age 4
Cough	Occurrence of cough at age 4
Nasal symptoms	Occurrence of nasal symptoms at age 4
Nocturnal symptoms	Occurrence of nocturnal asthma symptoms at age 4
Eczema	Eczema status at age 4
Hay fever	Hay fever status at age 4
Atopy	Atopy status (sensitisation to one or more allergens) at age 4
Monosensitisation	Sensitisation to one allergen at age 4
Polysensitisation	Sensitisation to two or more allergens at age 4
Parental smoking	Household parental smoking status at age 4
Dog	Household pet dog at age 4
Cat	Household pet cat at age 4
Furry pet	Household furry pet (dog, cat or other animal) at age 4

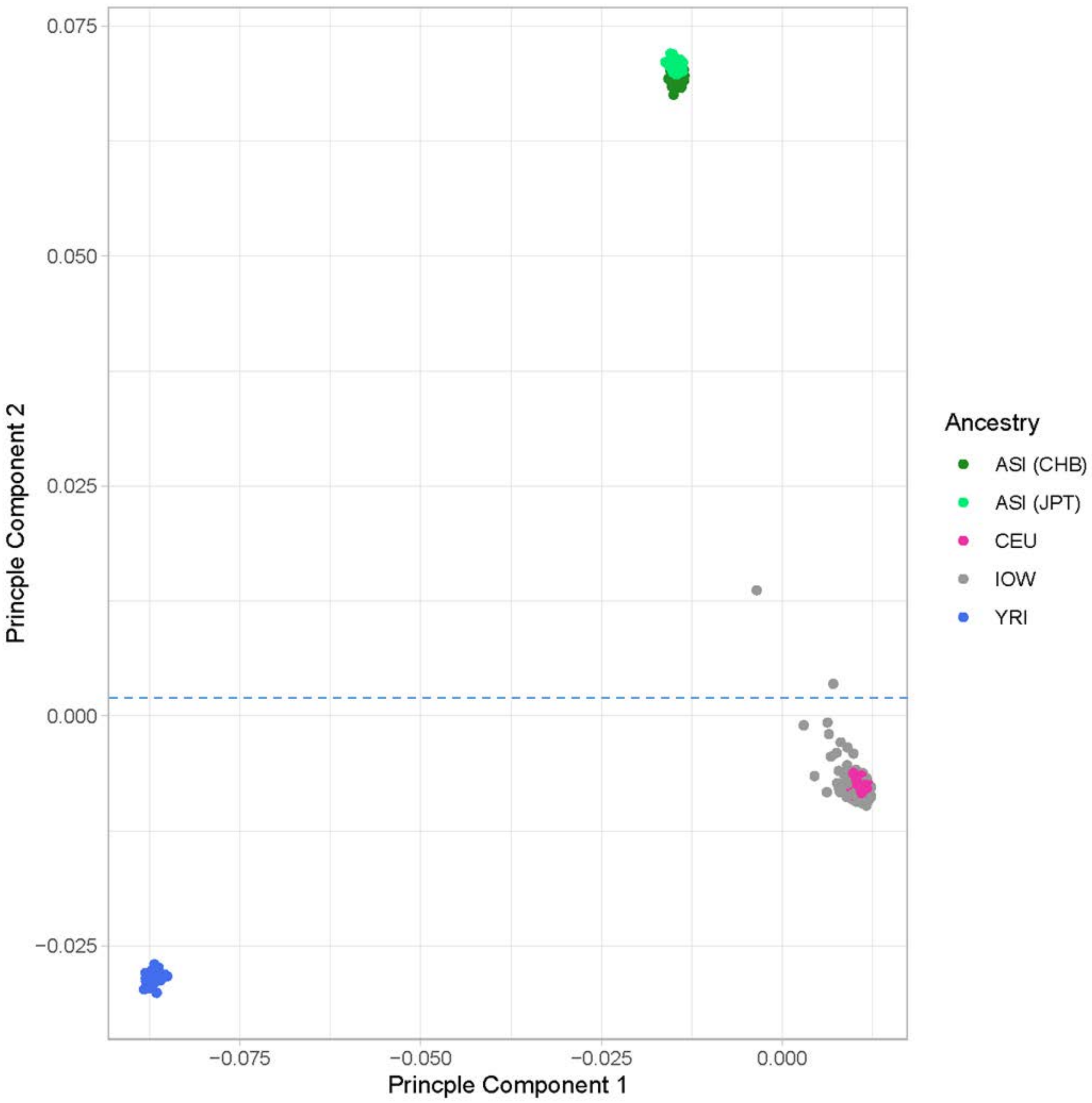


Figure A1 Visualisation of the first two principal components to identify the ancestry of individuals in the IOWBC

Population structure was assessed by principal component analysis (PCA), comparing the IOWBC (grey) with the European descent (pink), Yoruba (blue), Hans Chinese (dark green) and Japanese (light green) HapMap3 reference populations. Non-European individuals in the IOWBC were excluded before analysis (blue dashed line indicates the exclusion threshold considered).

Table A2 Hyperparameters tuned during the development of each machine learning algorithm

Algorithm	Hyperparameters	Description	Search range
Support Vector Machine	Cost	Regularisation term	100 values between 10^{-3} and 10^{2} [a]
	Gamma	Scalar term for the RBF and polynomial kernels	100 values between 10^{-2} and 10^{2} [a]
	Degree	Degree term for the polynomial kernel	1,2,3,…,10
Decision Tree	Max tree depth	The maximum depth each tree should be constructed to	1,2,3,…,32 or None
	Min samples split	The minimum number of samples needed to split a node	2,3,4,…,11
	Max features	The maximum number of features to consider to find the best split	'log2', 'sqrt', None
	Splitter	Criteria used to choose the split at a node	'best', 'random'
	Criterion	Criteria used to determine the quality of a node split	Gini, entropy
Random forest	N estimators (trees)	The number of trees used to construct the forest	1,2,4,8,16,32,64,100,200
	Max tree depth	The maximum depth each tree should be constructed to	1,2,3,…,32
	Min samples split	The minimum number of samples needed to split a node	2,3,4,…,11
	Max features	The maximum number of features to consider to find the best split	'log2', 'sqrt', None
	Criterion	Criteria used to determine the quality of a node split	Gini, entropy
	Bootstrap	Determines whether bootstrapping with replacement should be used to build the trees	True, False

Multilayer Perceptron	Hidden layers	The number of neurons in each hidden layer	(1,),(2,),…(11,) (1,1),(2,2),…(11,11)[b]
	Activation	The activation function for the hidden layers	'relu', 'identity', 'tanh', 'logistic’
	Solver	Criteria used to optimise the weights of the connections	'lbfgs', 'sgd', 'adam'
	Alpha	Regularisation term	10^{-1} ,10^{-2}, 10^{-6}
	Learning rate	The rate at which to update the weights	'constant', 'invscaling', 'adaptive'
	Initial learning rate	The initial learning rate	0.1,0.2,...,0.9
KNN	Number of neighbours (k)	The number of neighbours	1,2,3,…,100
	Weight	Determines whether each neighbour should be weighted equally or based on their distance	Uniform, distance
	Power	Specifies the distance measure to use	Manhattan, Euclidean
Naïve Bayes	Distribution	Determines which distribution each feature is assumed to follow	Continuous features = Gaussian distribution. Categorical features= multinomial distribution

[a] Specifies the parameter space for the random search strategy. Based on the results of the random search, a refined grid search across 500 steps was specified.

[b] Number of neurons in each hidden layer, where (1,) represents 1 neuron in the first hidden layer, with no further hidden layers; and (1,1) represents 1 neuron in the first hidden layer and 1 in the second hidden layer.

[c] The naïve Bayes algorithm did not undergo any hyperparameter search. Rather than being a hyperparameter to tune, the distributions were specified for each variable type at the time of model development.

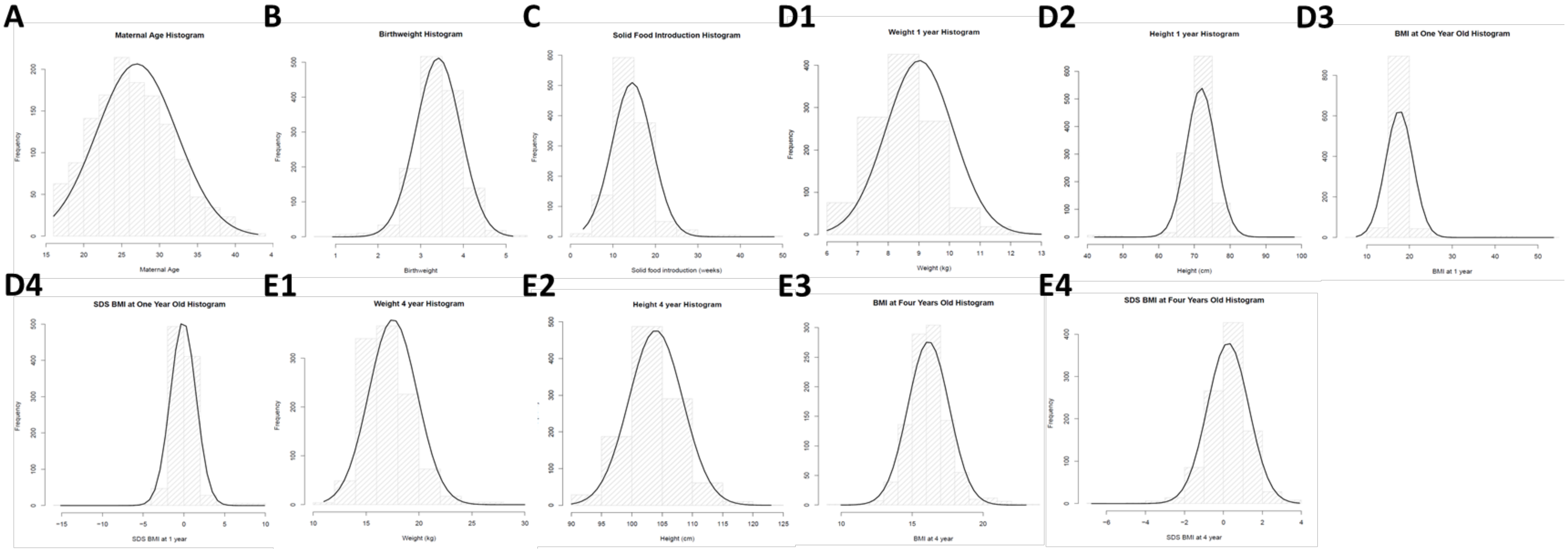


Figure A2 Data distribution of the continuous candidate features considered during the development of the early life and preschool models

Table A3 Comparability between predictor and outcome definitions in the IOWBC and MAAS

Variable	IOWBC definition	MAAS definition [a]
Maternal age	Maternal age at booking	Maternal age at birth of child
Birthweight	Birth weight (kg)	Birth weight (kg)
Total breastfeeding	Total breastfeeding duration	Breast feeding duration
Age of solid food introduction	Age of introduction of cereals/solids (weeks)	At what age did your baby begin solid foods? (weeks)
Early life BMI	BMI at age 1, standardised against the British 1990 growth reference	BMI at age 1, standardised against the British 1990 growth reference
Early life wheeze	Frequency of asthma wheezing episodes at either 1 or 2 years Categorised as no wheeze, occasional (1-3 times per year), frequent (12+ times per year)	Has or does your child's chest ever wheeze or whistle? If answer was yes, what best describes your child's wheezing (at either 1 or 3 years)? Categorised as no wheeze, 1-2 times or from time to time (occasional), every day (frequent)
Early life cough	Asthmatic cough at either 1 or 2 years	Does your child usually have a cough apart from with colds at 1 or 3 years
Preschool BMI	SDS BMI at age 4	SDS BMI at age 5
Preschool wheeze	Frequency of wheezing at 4YR	Current wheeze age 5 years
Preschool cough	Any asthmatic cough at 4 YR	Does your child usually have a cough during the day apart from with colds?

Preschool nocturnal symptoms	Any nocturnal symptoms at 4YR	Does your child usually have a cough at night apart from with colds? Or, in the last 12 months how often - on average - has your child's sleep been disturbed by wheezing
Preschool atopy	Sensitisation to one or more allergens at age 4 Tested allergens included HDM, milk, egg, cat, dog, grass, wheat, soya, peanut, cod, Cladosporium, Alternaria	Sensitisation to one or more allergens at age 5 Tested allergens included HDM, cat, dog, pollen, mould, milk, egg
Preschool polysensitisation	Sensitisation to two or more allergens by age 4 Tested allergens: HDM, milk, egg, cat, dog, grass, wheat, soya, peanut, cod, Cladosporium, Alternaria	Sensitisation to two or more allergens by age 5 Tested allergens included HDM, cat, dog, pollen, mould, milk, egg
Maternal socioeconomic status	Maternal socioeconomic status based on household income, number of rooms in the house and maternal education level Categorised into 5 groups: very low, low, low-middle, middle and high.	Maternal socioeconomic status based on professional occupation Categorised into 4 groups: routine (low), intermediate (low-middle), managerial (middle) and professional (high).
School-age asthma	Doctor diagnose asthma plus wheeze in the last 12 month and/or asthma treatment, evaluated at age 10	Doctor diagnose asthma plus wheeze in the last 12 month and/or asthma treatment, evaluated at ages 8 and 11

[a] MAAS variable categorisations are given as: categorisation of the MAAS variables (IOWBC equivalent categorisation).

Sensitisation to allergens identified by positive skin prick tests to allergens

HDM=house dust mite; SDS BMI= body mass index standardised against the British Growth Reference.

Table A4 Descriptive statistics for all candidate features considered for the development of the early life and preschool prediction models

	Total IOWBC (n=1368)		**Early life complete dataset (n=490)**		**Preschool complete dataset (n=373)**	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=70)**	**No asthma (n=420)**	**Asthma (n=55)**	**No asthma (n=318)**
Family history/ parent demographic predictors						
Maternal smoking at birth	47 (23.38)	276 (23.65)	15 (21.43)	75 (17.86)	12 (21.82)	61 (19.18)
Paternal smoking at birth	79 (39.30)	440 (37.70)	24 (34.29)	137 (32.62)	21 (38.18)	112 (35.22)
Maternal asthma	29 (14.43)	113 (9.68)	5 (7.14)	39 (9.29)	3 (5.45)	25 (7.86)
Maternal eczema	28 (13.93)	133 (11.40)	11 (15.71)	49 (11.67)	10 (18.18)	34 (10.69)
Maternal hay fever	50 (24.88)	219 (18.77)	13 (18.57)	85 (20.24)	12 (21.82)	62 (19.50)
Paternal asthma	27 (13.43)	104 (8.91)	10 (14.29)	33 (7.86)	8 (14.55)	27 (8.49)
Paternal eczema	19 (9.45)	70 (6.00)	10 (14.29)*	25 (5.95)*	7 (12.73)	20 (6.29)
Paternal hay fever	35 (17.41)	166 (14.22)	14 (20.00)	54 (12.86)	13 (23.64)	40 (12.58)
Parity	95 (47.26)	573 (49.10)	42 (60.00)	240 (57.14)	36 (65.45)	180 (56.60)

	Total IOWBC (n=1368)		Early life complete dataset (n=490)		Preschool complete dataset (n=373)	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=70)**	**No asthma (n=420)**	**Asthma (n=55)**	**No asthma (n=318)**
SES						
Very low	25 (12.44)	163 (13.97)	-	-	10 (18.18)	30 (9.43)
Low	35 (17.41)	199 (17.05)	-	-	8 (14.55)	63 (19.81)
Low-middle	62 (30.85)	334 (28.62)	-	-	14 (25.45)	103 (32.39)
Middle	52 (26.37)	320 (27.42)	-	-	16 (29.09)	99 (31.13)
High	13 (6.47)	96 (8.23)	-	-	7 (12.73)	23 (7.23)
Perinatal/ at birth predictors						
Maternal age	201 (26.61, 5.44)	1167 (27.04, 5.26)	70 (27.44, 5.32)	420 (27.60, 4.91)	55 (27.98, 5.37)	318 (27.69, 4.95)
Prematurity						
Pre-term	9 (4.48)	32 (2.74)	1 (1.43)	7 (1.67)	1 (1.82)	4 (1.26)
Term	184 (91.54)	1103 (94.52)	67 (95.71)	411 (97.86)	53 (96.36)	312 (98.11)
Post-term	3 (1.49)	12 (1.03)	2 (2.86)	2 (0.48)	1 (1.82)	2 (0.63)
Caesarean delivery	18 (8.96)	86 (7.37)	8 (11.43)	39 (9.29)	7 (12.73)	31 (9.75)
Total breastfeeding						
Never	46 (22.89)	267 (22.88)	18 (25.71)	95 (22.62)	15 (27.27)	73 (22.96)
<3months	66 (32.84)	352 (30.16)	29 (41.43)	137 (32.62)	22 (40.00)	105 (33.02)
3-6 months	22 (10.95)	164 (14.05)	4 (5.71)	69 (16.43)	3 (5.45)	56 (17.61)
>6 months	37 (18.41)	264 (22.62)	19 (27.14)	119 (28.33)	15 (27.27)	84 (26.42)

	Total IOWBC (n=1368)		Early life complete dataset (n=490)		Preschool complete dataset (n=373)	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=70)**	**No asthma (n=420)**	**Asthma (n=55)**	**No asthma (n=318)**
Exclusive breastfeeding						
Never	55 (27.36)	334 (28.62)	25 (35.71)	125 (29.76)	21 (38.18)	91 (28.62)
<3 months	85 (42.29)	489 (41.90)	35 (50.00)	191 (45.48)	27 (49.09)	146 (45.91)
>3 months	31 (15.42)	224 (19.19)	10 (14.29)	104 (24.76)	7 (12.72)	81 (25.47)
Age of solid food introduction	168 (14.36, 4.51)	1026 (14.34, 4.12)	70 (13.96, 4.08)	420 (14.45, 4.08)	55 (13.96, 4.24)	318 (14.59, 4.07)
Birthweight	199 (3.34, 0.52)*	1142 (3.44, 0.50)*	70 (3.45, 0.51)	420 (3.47, 0.52)	55 (3.48, 0.56)	318 (3.45, 0.49)
Sex	*					
Male	118 (58.71)	578 (49.53)	40 (57.14)	191 (45.48)	31 (56.36)	143 (44.97)
Female	83 (41.29)	589 (50.47)	30 (42.86)	229 (54.52)	24 (43.64)	175 (55.03)
Season of birth						
Autumn	38 (18.91)	243 (20.82)	13 (18.57)	101 (24.05)	12 (21.82)	86 (27.04)
Winter	64 (31.84)	382 (32.73)	19 (27.14)	117 (27.86)	14 (25.45)	77 (24.21)
Spring	51 (25.37)	274 (23.48)	19 (27.14)	100 (23.81)	15 (27.27)	80 (25.16)
Summer	48 (23.88)	268 (22.96)	19 (27.14)	102 (24.29)	14 (25.45)	75 (23.58)
Dog	51 (25.37)	346 (29.65)	15 (21.43)	119 (28.33)	14 (25.45)	92 (28.93)
Cat	57 (28.36)	397 (34.02)	24 (34.29)	146 (34.76)	20 (36.36)	113 (35.53)
Furry pet	95 (47.26)	636 (54.50)	31 (44.29)	226 (53.81)	26 (47.27)	175 (55.03)

	Total IOWBC (n=1368)		Early life complete dataset (n=490)		Preschool complete dataset (n=373)	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=70)**	**No asthma (n=420)**	**Asthma (n=55)**	**No asthma (n=318)**
Early life Predictors						
BMI at 1YR	135 (-0.15, 1.15)	851 (-0.16, 1.22)	70 (-0.13, 1.12)	420 (-0.17, 1.19)	55 (-0.20, 1.13)	318 (-0.18, 1.19)
Wheeze	*	*	*	*	*	*
Never	78 (38.81)	739 (63.32)	39 (55.71)	341 (81.19)	30 (54.55)	259 (81.45)
Occasional	14 (6.97)	63 (5.40)	6 (8.57)	28 (6.67)	5 (9.09)	20 (6.29)
Frequent	56 (27.86)	124 (10.63)	25 (35.71)	51 (12.14)	20 (36.36)	39 (12.26)
Wheeze without cold	56 (27.86)*	124 (10.63)*	25 (35.71)*	51 (12.14)*	20 (36.36)*	39 (12.26)*
Cough	70 (34.83)*	174 (14.91)*	31 (44.29)*	74 (17.62)*	25 (45.45)*	55 (17.30)*
Nasal symptoms	60 (29.85)*	239 (20.48)*	25 (35.71)*	89 (21.19)*	19 (34.55)*	66 (20.75)*
Chest infection	54 (26.87)*	139 (11.91)*	19 (27.14)*	45 (10.71)*	15 (27.27)*	34 (10.69)*
Nocturnal symptoms	68 (33.83)*	161 (13.80)*	29 (41.43)*	73 (17.28)*	23 (41.82)*	54 (16.98)*
Eczema	66 (32.84)*	247 (21.17)*	24 (34.29)*	100 (23.81)*	20 (36.36)	72 (22.64)
Hay fever	36 (17.91)	165 (14.14)	17 (24.29)	67 (15.95)	13 (23.64)	50 (15.72)
Atopy	44 (21.89)*	58 (4.97)*	17 (24.29)*	28 (6.67)*	12 (21.82)*	14 (4.40)*
Monosensitisation	37 (18.41)*	51 (4.37)*	15 (21.43)*	26 (6.19)*	11 (20.00)*	13 (4.09)*
Polysensitisation	10 (4.98)	9 (0.77)	3 (4.29)	3 (0.71)	2 (3.64)	1 (0.31)
Parental smoking						
Never	69 (34.33)	470 (40.27)	33 (47.14)	231 (55.00)	24 (43.64)	168 (52.83)
Ex-smoker	5 (2.49)	54 (4.63)	2 (2.86)	29 (6.90)	2 (3.64)	23 (7.23)
Current	93 (46.27)	488 (41.82)	35 (50.00)	160 (38.10)	29 (52.73)	127 (39.94)
Dog	41 (20.40)	327 (28.02)	15 (21.43)	126 (30.00)	14 (25.45)	99 (31.13)

	Total IOWBC (n=1368)		**Early life complete dataset (n=490)**		**Preschool complete dataset (n=373)**	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=70)**	**No asthma (n=420)**	**Asthma (n=55)**	**No asthma (n=318)**
Cat	131 (65.17)	821 (70.35)	59 (84.29)	349 (83.10)	46 (83.64)	264 (83.02)
Furry pet	155 (77.11)	1001 (85.78)	66 (94.29)	412 (98.10)	52 (94.55)	311 (97.80)
Residence on a farm	6 (2.99)	43 (3.68)	4 (5.71)	22 (5.24)	3 (5.45)	13 (4.09)
Preschool predictors						
SDS BMI	146 (0.21, 1.03)	855 (0.23, 1.04)			55 (0.28, 0.88)	318 (0.28, 0.95)
Wheeze	*	*			*	*
Never	85 (42.29)	879 (75.32)			28 (50.91)	281 (88.36)
Occasional	18 (8.96)	34 (2.91)			7 (12.73)	10 (3.14)
Frequent	70 (34.83)	75 (6.43)			20 (36.36)	27 (8.49)
Wheeze without cold	70 (34.83)*	75 (6.43)*			20 (36.36)*	27 (8.49)*
Cough	99 (49.25)*	128 (10.97)*			32 (58.18)*	39 (12.26)*
Nasal symptoms	61 (30.35)*	136 (11.65)*			21 (38.18)*	43 (13.52)*
Nocturnal symptoms	94 (46.77)*	128 (10.97)*	-	-	30 (54.55)*	41 (12.89)*
Eczema	47 (23.38)*	87 (7.46)*	-	-	10 (18.18)	28 (8.81)
Hay fever	29 (14.43)*	35 (3.00)*	-	-	10 (18.18)*	15 (4.72)*
Atopy	72 (35.82)*	94 (8.05)*	-	-	26 (47.27)*	43 (13.52)*
Monosensitisation	22 (10.95)*	53 (4.54)*	-	-	9 (16.36)	26 (8.18)
Polysensitisation	48 (23.88)*	38 (3.26)*	-	-	17 (30.91)*	17 (5.35)*

	Total IOWBC (n=1368)		Early life complete dataset (n=490)		Preschool complete dataset (n=373)	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=70)**	**No asthma (n=420)**	**Asthma (n=55)**	**No asthma (n=318)**
Parental smoking						
Never	62 (30.85)	424 (36.33)	-	-	22 (40.00)	157 (49.37)
Ex-smoker	19 (9.45)	122 (10.45)	-	-	6 (10.91)	47 (14.78)
Current	78 (38.81)	357 (30.59)	-	-	27 (49.09)	114 (35.85)
Dog	50 (24.88)	280 (23.99)	-	-	15 (27.27)	91 (28.62)
Cat	60 (29.85)	370 (31.71)	-	-	15 (47.27)	120 (37.74)
Furry pet	96 (47.76)	586 (50.21)	-	-	37 (67.27)	195 (61.32)

Summary data is reported as the number of individuals, with the mean and standard deviation ($\bar{x}$, s) for the continuous features of: maternal age, birthweight, age of solid food introduction, early life SDS BMI and preschool SDS BMI; or proportions for the remaining categorical features (%).

SDS BMI= body mass index standardised against the British Growth Reference.

*Statistically significant differences between asthmatic and non-asthmatic children at age 10 ($p<0.05$), assessed using an independent two sample t-test or Pearson’s Chi-square test for independence are identified for continuous and categorical features, respectively.

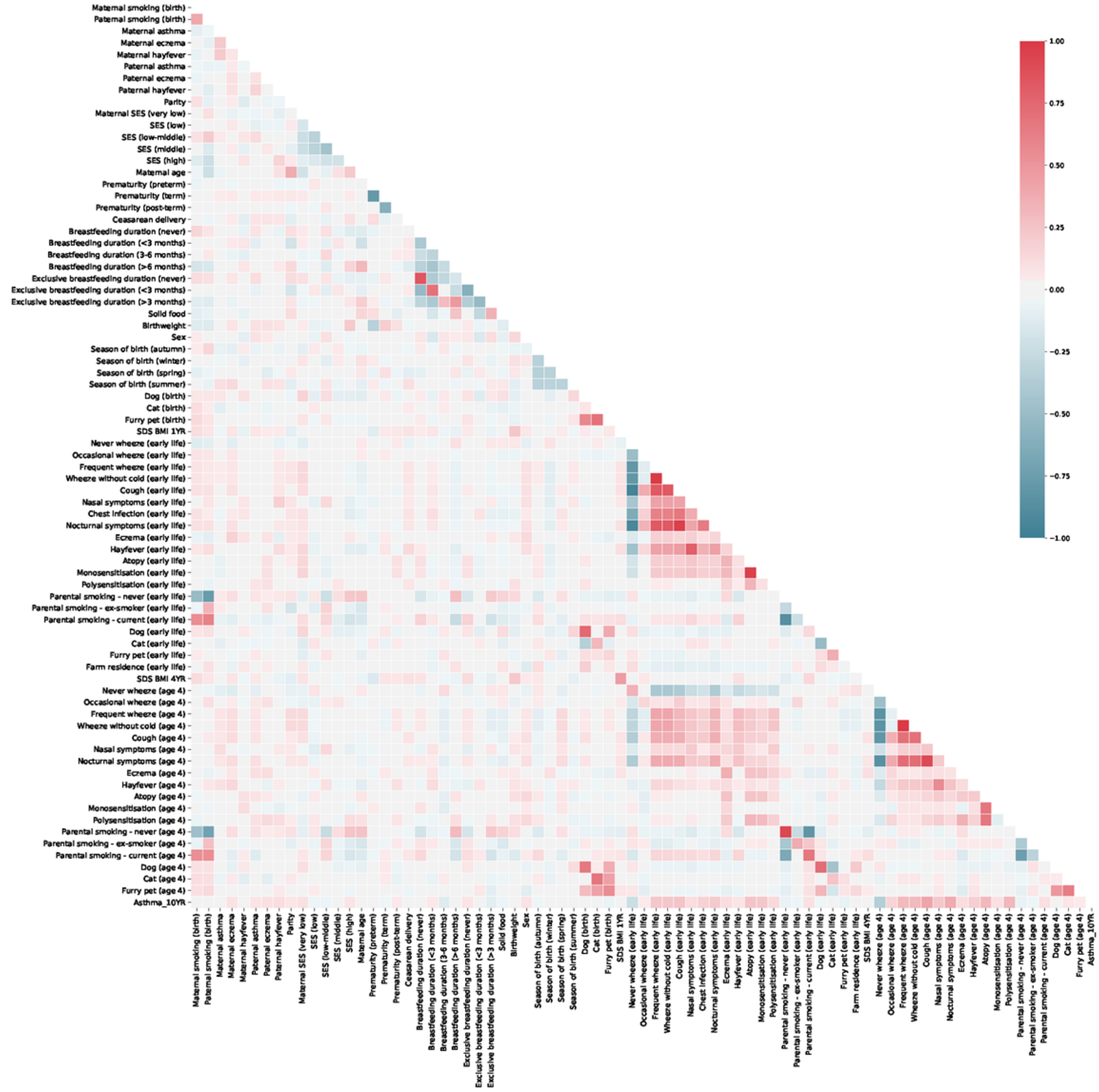


Figure A3 Correlation Matrix of the 54 candidate predictors considered for model development

Pearson's correlation coefficient was used to assess the collinearity between all pairs of candidate predictors. Correlation between predictors are visualised using the colour scale, with perfect positive correlations in red and perfect negative correlations in blue.

SDS BMI= body mass index standardised against the British Growth Reference.

Table A5 Top 20 candidate features for the early life model based on feature importance

Early life features: **Top 20**	**Feature importance** [a]	**Feature importance (RFECV)**[b]
SDS BMI 1YR	0.11	0.21
Birthweight	0.09	0.19
Maternal age	0.09	0.17
Age of solid food introduction	0.07	0.12
Total breastfeeding duration	0.04	0.08
Early life wheeze	0.04	0.08
Maternal socioeconomic status	0.04	0.08
Early life cough	0.03	0.07
Exclusive breastfeeding duration	0.03	-
Early life monosensitisation	0.02	-
Early life atopy	0.02	-
Sex	0.02	-
Pet cat at birth	0.02	-
Early life chest infection	0.02	-
Early life parental smoking	0.02	-
Early life nocturnal symptoms	0.02	-
Furry pet at birth	0.02	-
Early life wheeze without cold	0.02	-
Early life eczema	0.02	-
Parity	0.02	-

[a] Feature importance calculated, based on the Gini impurity, using all 39 candidate features considered.

[b] Feature importance calculated, based on the Gini impurity, using only the subset of features identified through the RFECV feature selection process.

SDS BMI= body mass index standardised against the British Growth Reference.

Table A6 Top 20 candidate features for the preschool model based on feature importance

Preschool features: Top 20	**Feature importance [a]**	**Feature importance (RFECV)[b]**
Preschool atopy	0.06	0.09
SDS BMI 1YR	0.06	0.12
Preschool cough	0.06	0.12
SDS BMI 4YR	0.05	0.10
Maternal age	0.05	0.12
Preschool nocturnal symptoms	0.05	0.07
Birthweight	0.05	0.09
Preschool wheeze	0.04	0.06
Age of solid food introduction	0.03	0.07
Preschool wheeze without cold	0.03	-
Preschool polysensitisation	0.03	0.06
Preschool nasal symptoms	0.03	-
Exclusive breastfeeding duration	0.03	-
Maternal socioeconomic status	0.03	0.06
Early life cough	0.02	-
Preschool parental smoking	0.02	-
Total breastfeeding duration	0.02	0.05
Paternal hay fever	0.02	-
Early life wheeze	0.02	-
Early life parental smoking	0.01	-

[a] Feature importance calculated, based on the Gini impurity, using all 54 candidate features considered.

[b] Feature importance calculated, based on the Gini impurity, using only the subset of features identified through the RFECV feature selection process.

SDS BMI= body mass index standardised against the British Growth Reference.

Table A7 Description of the preschool datasets following oversampling

	Sample size	Number of cases	Number of controls
Preschool dataset	548	76	472
Preschool training set	365	51	314
Oversampled by 25%	378	64	314
Oversampled by 50%	391	77	314
Oversampled by 100%	416	102	314
Oversampled by 150%	442	128	314
Oversampled by 200%	467	153	314
Oversampled by 250%	493	179	314
Oversampled by 300%	518	204	314
Preschool validation set	183	25	158

Table A8 Performance of the preschool naïve Bayes model following oversampling

	Accuracy	BA	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Initial Model	0.860	0.780	0.833	0.667	0.892	0.5	0943	6.157	0.374	0.572	280, 34, 17, 34
	0.825	0.764	0.679	0.680	0.848	0.415	0.944	4.48	0.377	0.515	134, 24, 8, 17
Oversampled cases 25%	0.833	0.763	0.803	0.656	0.869	0.506	0.925	5.026	0.395	0.571	273,41, 22,42
	0.809	0.772	0.657	0.720	0.823	0.391	0.949	4.063	0.340	0.507	130, 28, 7,18
Oversampled cases 50%	0.821	0.756	0.803	0.649	0.863	0.538	0.909	4.742	0.406	0.588	271, 43, 27, 50
	0.803	0.768	0.647	0.720	0.816	0.383	0.949	3.923	0.343	0.500	129, 29, 7, 18
Oversampled cases 100%	0.800	0.752	0.796	0.657	0.847	0.583	0.884	4.297	0.406	0.618	266, 48, 35, 67
	0.792	0.762	0.637	0.720	0.804	0.367	0.948	3.670	0.348	0.486	127, 31, 7, 18
Oversampled cases 150%	0.801	0.763	0.818	0.672	0.854	0.652	0.865	4.586	0.384	0.662	268, 46, 42, 86
	0.792	0.762	0.641	0.720	0.804	0.367	0.948	3.670	0.348	0.486	127, 31, 7, 18

	Accuracy	BA	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 Score	TN, FP, FN, TP [a]
Oversampled cases 200%	0.799	0.772	0.809	0.693	0.850	0.693	0.850	4.629	0.361	0.693	267, 47, 47, 106
	0.792	0.762	0.635	0.720	0.804	0.367	0.948	3.670	0.348	0.486	127, 31, 7, 18
Oversampled cases 250%	0.793	0.773	0.814	0.698	0.847	0.723	0.831	4.568	0.356	0.710	266, 48, 54, 125
	0.792	0.762	0.606	0.720	0.804	0.367	0.948	3.670	0.348	0.486	127, 31, 7, 18
Oversampled cases 300%	0.792	0.777	0.829	0.706	0.847	0.750	0.816	4.618	0.347	0.727	266, 48, 60, 144
	0.792	0.762	0.636	0.720	0.804	0.367	0.948	3.670	0.348	0.486	127, 31, 7, 18

Shaded rows report performance on the training set, unshaded rows report performance on the holdout validation set.

[a] The final row presents the confusion matrix for the model classifications, where TN=true negatives, FP=false positives, FN=false negatives, TP=true positives.

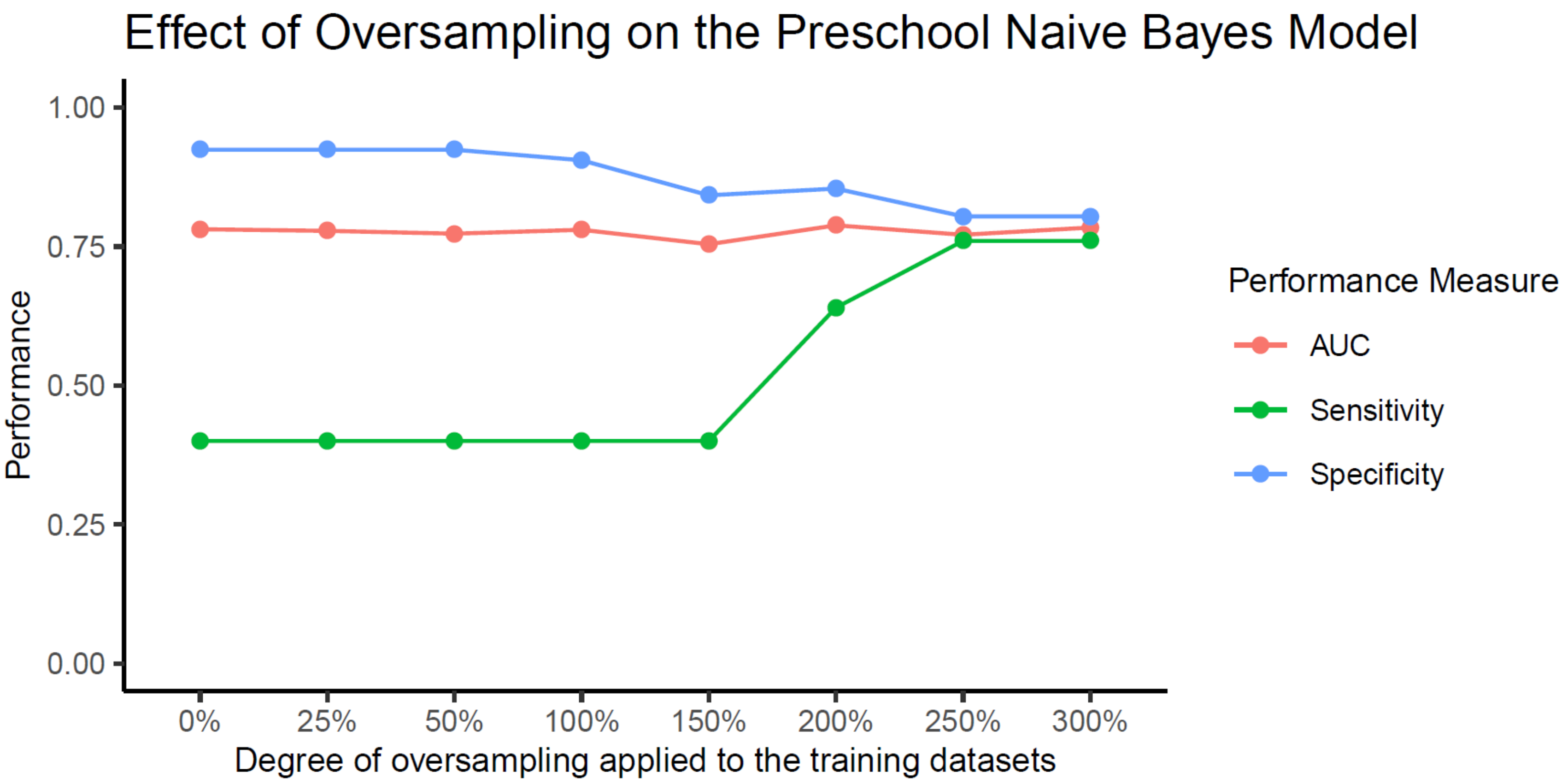


Figure A4 Effect of oversampling on the performance of the preschool naive Bayes model

Table A9 Distribution of CAPE and CAPP model predictors for individuals in the IOWBC and MAAS at each asthma prediction time-point

	Total IOWBC (n=1368)		**MAAS 8YR (n=1018)**		**MAAS 11YR(n=898)**	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=144)**	**No asthma (n=874)**	**Asthma (n=116)**	**No asthma (n=782)**
Maternal age	201 (26.61, 5.44)	1167 (27.04, 5.26)	116 (30.53, 5.09)	842 (20.66, 4.67)	94 (29.59, 4.94))*	762 (30.88, 4.61)*
Birthweight	199 (3.34, 0.52)*	1142 (3.44, 0.50)*	132 (3.44, 0.50)	845 (3.49, 0.49)	107 (3.41, 0.51)	757 (3.49, 0.49)
Age of solid food introduction	168 (14.36, 4.51)	1026 (14.34, 4.12)	51 (14.88, 3.83)	392 (14.67, 3.52)	44 (14.93, 5.03)	351 (14.69, 3.34)
Breastfeeding duration						
Never	46 (22.89)	267 (22.88)	47 (23.64)	281 (32.15)	32 (27.59)	236 (30.18)
<3months	66 (32.84)	352 (30.16)	33 (22.92)	214 (24.49)	30 (25.86)	190 (24.30)
3-6 months	22 (10.95)	164 (14.05)	24 (16.67)	162 (18.54)	16 (12.79)	157 (20.08)
>6 months	37 (18.41)	264 (22.62)	22 (15.28)	194 (22.20)	23 (19.83)	181 (23.15)
Early life SDS BMI	135 (-0.15, 1.15)	851 (-0.16, 1.22)	49 (-0.18, 1.00)	387 (-0.25, 1.11)	43 (-0.04, 1.09)	347 (-0.32, 1.12)
Preschool SDS BMI	146 (0.21, 1.03)	855 (0.23, 1.04)	134 (0.57, 0.95)	804 (0.46, 0.94)	113 (0.65, 0.90)*	731 (0.42, 0.94)*
Early life wheeze	*	*	*	*	*	*
Never	78 (38.81)	739 (63.32)	8 (5.56)	212 (24.26)	9 (7.76)	189 (24.17)
Occasional	14 (6.97)	63 (5.40)	93 (64.58)	339 (38.79)	67 (57.76)	299 (38.24)
Frequent	56 (27.86)	124 (10.63)	4 (2.78)	6 (0.69)	4 (3.45)	5 (0.64)
Early life cough	*	*	*	*	*	*
No	78 (38.81)	749 (64.18)	31 (21.53)	318 (36.38)	28 (24.14)	281 (35.93)
Yes	70 (34.83)	174 (14.91)	34 (23.61)	135 (15.45)	32 (27.59) *	117 (14.96) *

	Total IOWBC (n=1368)		MAAS 8YR (n=1018)		MAAS 11YR(n=898)	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=144)**	**No asthma (n=874)**	**Asthma (n=116)**	**No asthma (n=782)**
Preschool wheeze	*	*	*	*	*	*
Never	85 (42.29)	879 (75.32)	44 (30.56)	721 (82.49)	40 (34.48)	655 (83.76)
Occasional	18 (8.96)	34 (2.91)	89 (61.81)	116 (13.27)	66 (56.90)	105 (13.43)
Frequent	70 (34.83)	75 (6.43)	5 (3.47)	6 (0.69)	6 (5.17)	3 (0.38)
Preschool cough	*	*	*	*	*	*
No	74 (36.82)	860 (73.69)	70 (48.61)	711 (81.35)	62 (53.45)	642 (82.10)
Yes	99 (49.25)	128 (10.97)	68 (47.22)	132 (15.10)	50 (43.10)	121 (15.47)
Preschool nocturnal symptoms	*	*	*	*	*	*
No	79 (39.30)	860 (73.69)	49 (34.03)	700 (80.09)	47 (40.52)	629 (80.43)
Yes	94 (46.77)	128 (10.97)	89 (61.81)	143 (16.36)	65 (56.03)	134 (17.14)
Preschool atopy status	*	*	*	*	*	*
No	67 (33.33)	670 (57.41)	52 (36.11)	573 (65.56)	38 (32.76)	530 (67.77)
Yes	72 (35.82)	94 (8.05)	76 (52.78)	197 (22.54)	69 (59.48)	169 (21.61)
Preschool polysensitisation status	*	*	*	*	*	*
No	93 (46.27)	874 (74.89)	77 (53.47)	675 (77.23)	62 (53.45)	612 (78.26)
Yes	48 (23.88)	38 (3.26)	47 (32.64)	77 (8.81)	41 (35.34)	73 (9.34)

	Total IOWBC (n=1368)		MAAS 8YR (n=1018)		MAAS 11YR(n=898)	
	Asthma (n=201)	**No asthma (n=1167)**	**Asthma (n=144)**	**No asthma (n=874)**	**Asthma (n=116)**	**No asthma (n=782)**
Maternal socioeconomic status						
Very low	25 (12.44)	163 (13.97)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Low	35 (17.41)	199 (17.05)	11 (7.64)	66 (7.55)	9 (7.76)	54 (6.91)
Low-Mid	62 (30.85)	334 (28.62)	15 (10.42)	137 (15.68)	10 (8.62)	129 (16.50)
Mid	52 (26.37)	320 (27.42)	41 (28.47)	388 (44.39)	30 (25.86)	375 (47.95)
High	13 (6.47)	96 (8.23)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)

The distribution of predictors is reported as the number of individuals, with mean and standard deviation ($\bar{x}$, s) for the continuous features of: maternal age, birthweight, age of solid food introduction, early life BMI and preschool BMI; or as proportions for the remaining categorical features (%). Where the number of individuals with data for a variable does not equal the total number of individuals detailed in the column, the difference indicates the number of individuals with missing data.

SDS BMI= body mass index standardised against the British Growth Reference.

* Statistically significant difference between asthmatic and non-asthmatic children ($p<0.05$), assessed using an independent two sample t-test or Pearson's Chi-square test for independence for continuous and categorical features, respectively.

Table A10 Performance of the CAPE model using the original and SHAP restricted features

Model	Dataset (sample size)	Balanced accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 score
CAPE – 8 features	IOWBC test (n=255)	0.71 (0.62-0.78)	0.71 (0.61-0.80)	0.74 (0.56-0.88)	0.68 (0.62-0.74)	0.26 (0.21-0.32)	0.94 (0.91-0.97)	2.29 (1.69-3.01)	0.39 (0.18-0.63)	0.38 (0.31-0.46)
	MAAS 8YR (n=322)	0.67 (0.60-0.74)	0.71 (0.63-0.79)	0.84 (0.71-0.95)	0.51 (0.45-0.56)	0.19 (0.16-0.21)	0.96 (0.93-0.99)	1.71 (1.40-2.03)	0.31 (0.10-0.57)	0.30 (0.26-0.35)
	MAAS 11YR (n=299)	0.68 (0.60-0.74)	0.71 (0.62-0.79)	0.84 (0.72-0.97)	0.51 (0.45-0.57)	0.17 (0.14-0.20)	0.96 (0.94-0.99)	1.72 (1.39-2.05)	0.31 (0.07-0.58)	0.28 (0.24-0.33)
CAPE – 2 features	IOWBC test (n=255)	0.63 (0.54-0.71)	0.61 (0.51-0.71)	0.38 (0.21-0.56)	0.87 (0.82-0.91)	0.31 (0.19-0.44)	0.90 (0.88-0.93)	2.91 (1.54-5.09)	0.71 (0.52-0.90)	0.34 (0.21-0.47)
	MAAS 8YR (n=502)	0.51 (0.49-0.54)	0.58 (0.49-0.67)	0.03 (0.00-0.08)	0.99 (0.98-1.00)	0.29 (0.00-1.00)	0.87 (0.87-0.88)	2.68 (0.00-10.00)	0.98 (0.93-1.02)	0.06 (0.00-0.14)
	MAAS 11YR (n=444)	0.52 (0.49-0.55)	0.58 (0.49-0.67)	0.05 (0.00-0.12)	0.99 (0.98-1.00)	0.43 (0.00-1.00)	0.87 (0.87-0.88)	4.99 (0.00-10.00)	0.96 (0.89-1.01)	0.09 (0.00-0.20)

Performance measures for each CAPE model were reported based on the threshold cut-off that maximised the Youden's Index (CAPE 8 features model threshold=0.42; CAPE 2 feature model threshold=0.68). Performance in MAAS was evaluated in all individuals with complete data for the predictors included in each model.

Table A 11 Performance of the CAPP model using the original and SHAP restricted features

Model	Dataset (sample size)	Balanced accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 score
CAPP – 12 features	IOWBC test (n=183)	0.80 (0.70-0.89)	0.82 (0.71-0.91)	0.72 (0.52-0.88)	0.88 (0.83-0.92)	0.47 (0.38-0.62)	0.95 (0.92-0.98)	5.99 (3.79-10.11)	0.32 (0.13-0.54)	0.56 (0.45-0.70)
	MAAS 8YR (n=282)	0.73 (0.64-0.81)	0.83 (0.75-0.90)	0.55 (0.36-0.70)	0.91 (0.88-0.95)	0.45 (0.33-0.59)	0.94 (0.92-0.96)	6.17 (3.64-10.69)	0.50 (0.33-0.69)	0.49 (0.36-0.62)
	MAAS 11YR (n=267)	0.70 (0.60-0.80)	0.80 (0.70-0.88)	0.50 (0.31-0.69)	0.90 (0.85-0.95)	0.46 (0.32-0.63)	0.91 (0.89-0.94)	5.07 (2.77-9.95)	0.55 (0.34-0.77)	0.48 (0.32-0.63)
CAPP – 3 features	IOWBC test (n=183)	0.78 (0.69-0.87)	0.80 (0.70-0.89)	0.76 (0.60-0.92)	0.80 (0.74-0.86)	0.38 (0.30-0.48)	0.85 (0.92-0.98)	3.87 (2.65-5.77)	0.30 (0.10-0.52)	0.51 (0.40-0.61)
	MAAS 8YR (n=872)	0.73 (0.68-0.77)	0.77 (0.73-0.81)	0.70 (0.62-0.78)	0.75 (0.72-0.78)	0.32 (0.28-0.35)	0.94 (0.92-0.95)	2.82 (2.40-3.30)	0.40 (0.29-0.51)	0.44 (0.39-0.48)
	MAAS 11YR (n=784)	0.72 (0.68-0.77)	0.78 (0.74-0.83)	0.70 (0.61-0.78)	0.75 (0.72-0.78)	0.30 (0.26-0.33)	0.94 (0.93-0.96)	2.81 (2.36-3.37)	0.40 (0.29-0.52)	0.42 (0.37-0.47)

Performance measures for each CAPP model were reported based on the threshold cut-off that maximised the Youden's Index (CAPP 12 features model threshold=0.73; CAPP 3 feature model threshold=0.74). Performance in MAAS was evaluated in all individuals with complete data for the predictors included in each model.

Table A12 Performance of the CAPP model with and without the predictors of sensitisation

	Dataset	Balanced Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	LR+	LR-	F_1 score
CAPP	IOWBC	0.80 (0.70-0.89)	0.82 (0.71-0.91)	0.72 (0.52-0.88)	0.88 (0.83-0.92)	0.47 (0.38-0.62)	0.95 (0.92-0.98)	5.99 (3.79-10.11)	0.32 (0.13-0.54)	0.56 (0.45-0.70)
	MAAS 8YR	0.73 (0.64-0.81)	0.83 (0.75-0.90)	0.55 (0.36-0.70)	0.91 (0.88-0.95)	0.45 (0.33-0.59)	0.94 (0.92-0.96)	6.17 (3.64-10.69)	0.50 (0.33-0.69)	0.49 (0.36-0.62)
	MAAS 11YR	0.73 (0.63-0.82)	0.79 (0.68-0.88)	0.55 (0.38-0.72)	0.90 (0.87-0.94)	0.41 (0.29-0.55)	0.94 (0.92-0.96)	5.71 (3.44-9.85)	0.50 (0.30-0.71)	0.47 (0.33-0.62)
CAPP - without sensitisation	IOWBC	0.75 (0.64-0.84)	0.72 (0.58-0.85)	0.64 (0.44-0.80)	0.85 (0.80-0.91)	0.41 (0.30-0.53)	0.94 (0.91-0.97)	4.40 (2.71-7.22)	0.42 (0.22-0.66)	0.50 (0.36-0.63)
	MAAS 8YR	0.67 (0.59-0.76)	0.79 (0.70-0.87)	0.47 (0.31-0.64)	0.87 (0.83-0.91)	0.33 (0.23-0.44)	0.92 (0.90-0.95)	3.69 (2.23-5.91)	0.61 (0.41-0.80)	0.39 (0.27-0.51)
	MAAS 11YR	0.65 (0.56-0.74)	0.70 (0.57-0.81)	0.42 (0.26-0.58)	0.87 (0.83-0.91)	0.29 (0.19-0.40)	0.92 (0.90-0.95)	3.32 (1.87-5.55)	0.66 (0.46-0.87)	0.34 (0.21-0.47)

Performance measures for each CAPP model were reported based on the threshold cut-off that maximised the Youden's Index (CAPP model threshold=0.73; CAPP without predictors of sensitisation threshold=0.45).

Table A13 Performance of all polygenic risk scores constructed using PRSice

P-value Threshold	**Goodness of fit (R$^{2)}$**	**AUC (95% CI)**	**No. SNPs in PRS**
5.00E-08	0.019201	0.5985 (0.5452-0.6518)	80
5.01E-05	0.023019	0.6044 (0.5515-0.6573)	93
0.0007	0.022516	0.6040 (0.5509-0.6571)	94
0.0011	0.023651	0.6059 (0.5529-0.6590)	95
0.0015	0.024237	0.6077 (0.5547-0.6606)	96
0.00205	0.025743	0.6105 (0.5578-0.6632)	97
0.0044	0.025789	0.6111 (0.5585-0.6637)	98
0.0061	0.025394	0.6108 (0.5583-0.6634)	99
0.0062	0.025838	0.6110 (0.5583-0.6636)	100
0.0066	0.025364	0.6098 (0.5571-0.6624)	101
0.0325	0.025747	0.6103 (0.5576-0.6630)	102
0.035	0.026158	0.6108 (0.5580-0.6635)	103
0.04515	0.026432	0.6113 (0.5585-0.6640)	104
0.04665	**0.026678**	**0.6119 (0.5591-0.6647)**	**105**
0.2248	0.026484	0.6117 (0.5589-0.6644)	106
0.2618	0.026558	0.6115 (0.5588-0.6643)	107
0.2656	0.026389	0.6112 (0.5584-0.6640)	108
0.3233	0.026179	0.6110 (0.5583-0.6638)	109
0.4214	0.026207	0. 6111 (0.5582-0.6639)	110
0.6275	0.026280	0.6112 (0.5584-0.6640)	111
0.6543	0.026316	0.6111 (0.5584-0.6639)	112
0.7647	0.026371	0.6112 (0.5584-0.6640)	113
0.8332	0.026316	0.6112 (0.5584-0.6640)	114
0.852	0.026360	0.6112 (0.5584-0.6640)	115
0.9698	0.026354	0.6112 (0.5584-0.664)	116

The best PRS based on R^2 and AUC which was selected as the final childhood asthma PRS is highlighted in bold.

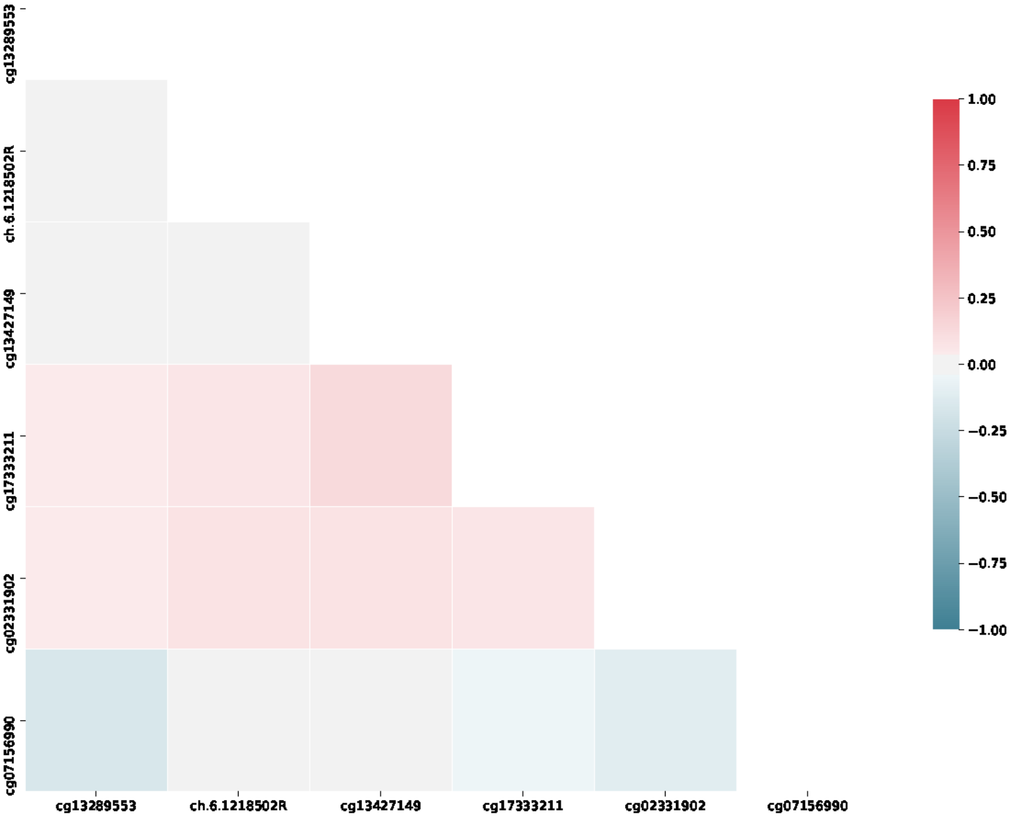


Figure A5 Correlation matrix of candidate CpGs considered for the newborn MRS

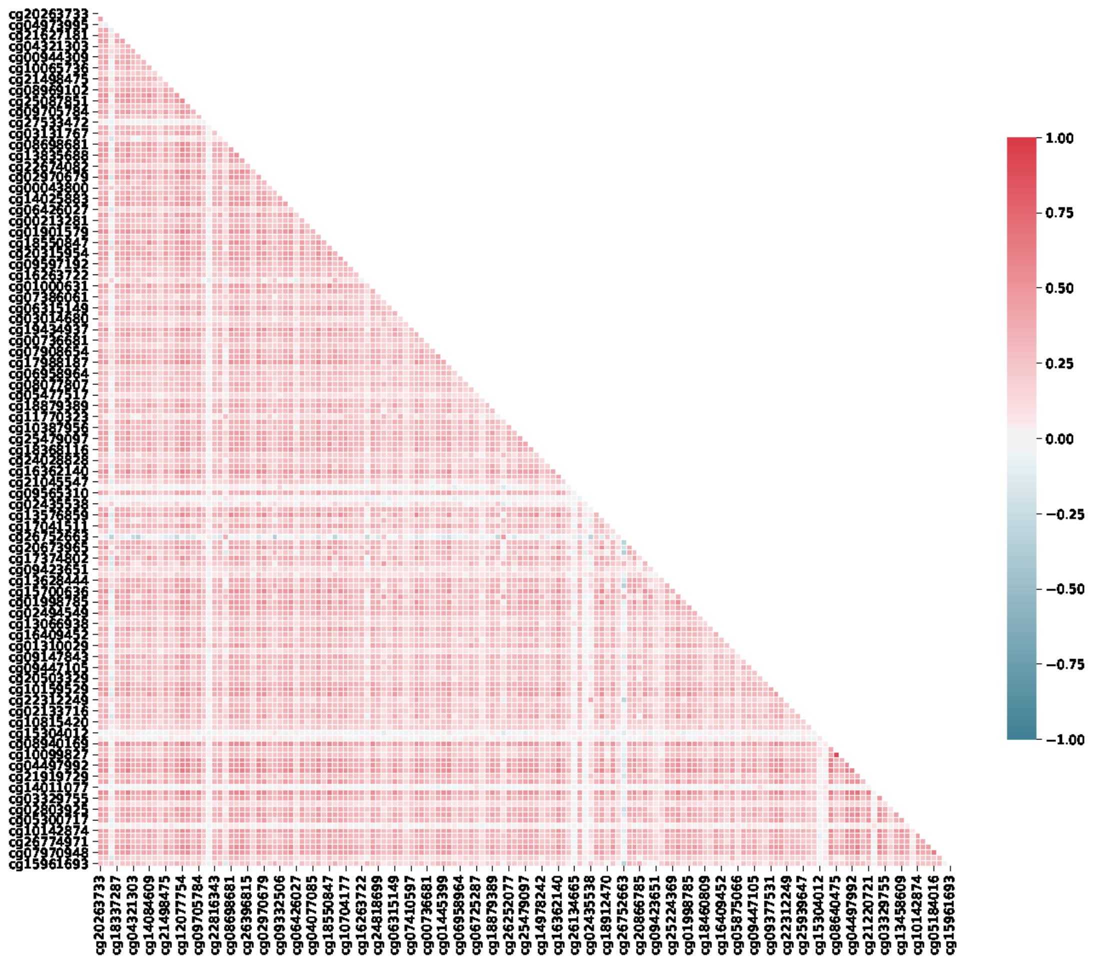


Figure A6 Correlation matrix of candidate CpGs considered for the childhood MRS

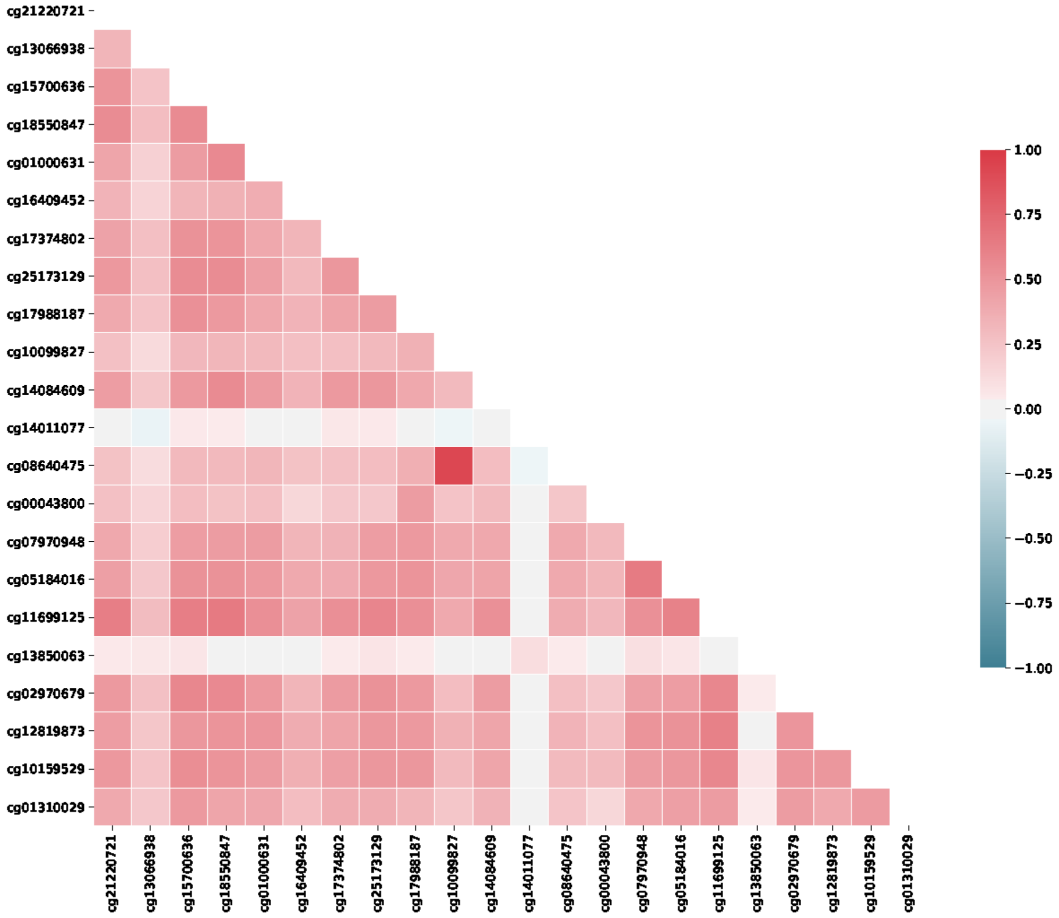


Figure A7 Correlation matrix of the 22 nearby CpGs considered for the childhood MRS

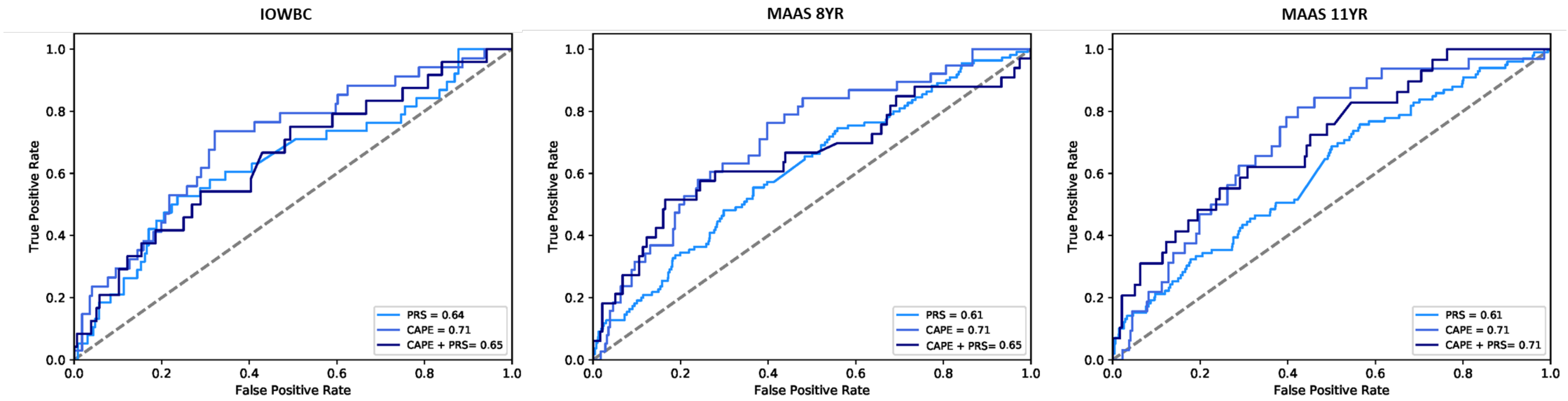


Figure A8 ROC curves comparing the performance of the CAPE model integrated with the PRS in the IOWBC and MAAS

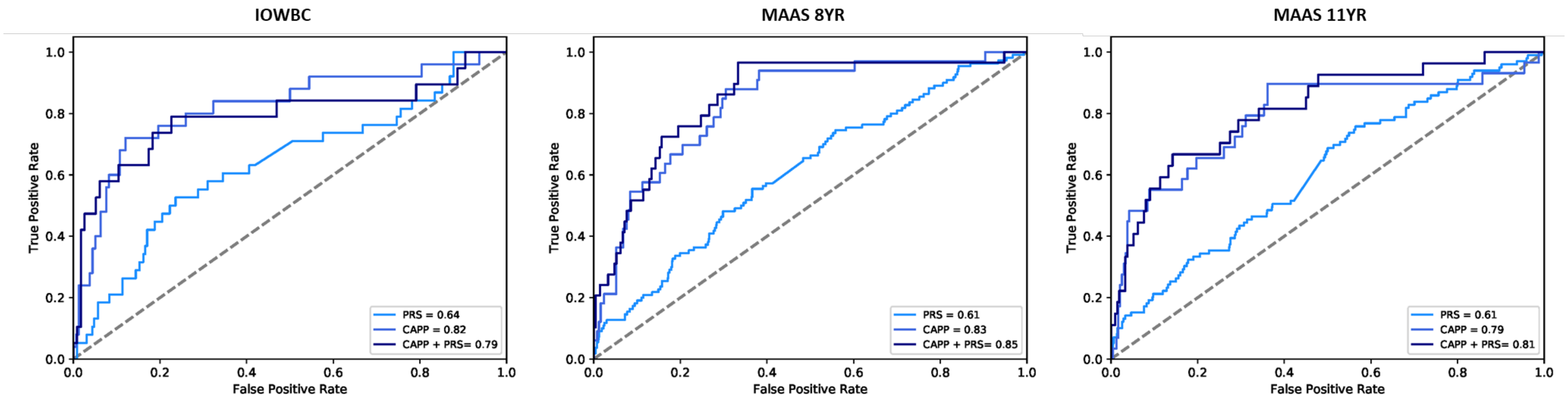


Figure A9 ROC curves comparing the performance of the CAPP model integrated with the PRS in the IOWBC and MAAS

Appendix B Systematic Review Database Searches

Table B1 Search strategy used for the Embase database search

Embase (1947 to 25th July 2019)
1. exp asthma/
2. asthma*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
3. wheezing/
4. wheez*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
5. 1 or 2 or 3 or 4
6. exp child/
7. (child or children or childhood or paediatric* or pediatric* or infant* or school-age or preschool or pre-school or early life or toddler*).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
8. 6 or 7
9. "prediction and forecasting"/ or prediction/ or computer prediction/
10. scoring system/
11. ((forecast* or predict* or risk*) adj3 (score* or model* or system* or formula* or value* or index* or tool*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
12. exp machine learning/
13. exp artificial intelligence/
14. intelligent system*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
15. 9 or 10 or 11 or 12 or 13 or 14
16. onset age/
17. (develop* or onset or outcome).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
18. 16 or 17
19. 5 and 18
20. 8 and 19
21. 15 and 20

Table B2 Search strategy used for the Medline database search

Medline Search Strategy (1946 to 25th July 2019)
1. exp Asthma/
2. asthma*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
3. wheez*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
4. 1 or 2 or 3
5. exp Child/
6. (child or children or childhood or paediatric* or pediatric* or infant* or school-age or preschool or pre-school or early life or toddler*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
7. exp Infant/
8. 5 or 6 or 7
9. ((forecast* or predict* or risk*) adj3 (score* or model* or system* or formula* or algorithm* or value* or index* or tool*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
10. exp Artificial Intelligence/
11. exp Machine Learning/
12. exp algorithms/
13. intelligent system*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
14. 9 or 10 or 11 or 12 or 13
15. "age of onset"/
16. (develop* or onset or outcome).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
17. 15 or 16
18. 4 and 17
19. 8 and 18
20. 14 and 19

Table B3 Search strategy used for the Web of Science database search

Web of Science Search Strategy
#1 TOPIC: (asthma*) OR TOPIC: (wheez*) DocType=All document types; Language=All languages;
#2 TOPIC: ((child OR children OR childhood OR paediatric* OR pediatric* OR infant* OR school-age OR preschool OR pre-school OR "early life" OR toddler*)) DocType=All document types; Language=All languages;
#3 TOPIC: (((forecast* OR predict* OR risk*) NEAR/3 (score* OR model* OR system* OR formula* OR algorithm OR value* OR index* or tool*))) DocType=All document types; Language=All languages;
#4 TOPIC: ("machine learning" OR "artificial intelligence" OR algorithm* OR "intelligent system*") DocType=All document types; Language=All languages;
#5 TOPIC: (develop* OR onset OR outcome*) DocType=All document types; Language=All languages;
#6 #4 OR #3 DocType=All document types; Language=All languages;
#7 #6 AND #2 AND #1 DocType=All document types; Language=All languages;
#8 #5 AND #1 DocType=All document types; Language=All languages;
#9 #8 AND #6 AND #2 DocType=All document types; Language=All languages;

Appendix C Code and Pseudocode

All scripts for data cleaning and reproducing analyses conducted in this thesis can be found at: doi.org/10.5258/SOTON/D1943. Permission to access data can be granted from David Hide Asthma & Allergy Research Centre (www.allergyresearch.org.uk/)

Pseudocode 1 Feature selection using Recursive Feature Elimination

Recursive Feature Elimination (RFE) with k-fold cross-validation

```
Input: training dataset containing p features and the outcome variable
Define k in k-fold cross-validation

Split the complete data into k partitions for cross-validation. Here k-1 of the k partitions will be used to train the model, and the final for validation
For each of the k cross-validation sets do
    For features p → 1 do
        For each training set do
            Train a balanced random forest classifier
            Assign weights of importance for each feature
            Rank the features by weighted importance
        End for
        Measure the performance on the validation set
        Remove the minimum ranking feature from the training dataset
    End for
    Average the performance across the validation sets
    Plot the number of features against the cross-validation prediction accuracy
End for
Identify the optimal number of features for the model
```

Pseudocode 2 Feature selection using Boruta

```
Boruta
Input: training dataset containing p features and the outcome variable

For n iterations of the random forest classifier do
    For feature space {1,2,…, p} do
        Replicate original feature
        Randomly shuffle each replicated feature across examples (generating a set of shadow features)
        Combine original and shadow features into an extended feature space
    End for
    Report the feature importance of all features
    Assign each original feature as a hit if the feature importance is higher than the maximal importance of the shadow features
End for
Count the number of iterations for which each feature was deemed important (hits)
Perform a two sided test of equality with the maximal importance amongst shadow variables where the expected number of hits for n iterations is E(n)=0.5n with standard deviation s=sqrt(0.25n)
Deem each feature important if the actual number of hits is significantly higher than expected, otherwise deem feature unimportant.
Identify the optimal subset of features for the model
```

List of References

1. Pavord ID, Beasley R, Agusti A, et al. After asthma: redefining airways diseases. The Lancet 2018;391:350-400.

2. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention2021.

3. Carr TF, Bleecker E. Asthma heterogeneity and severity. World Allergy Organization Journal 2016;9:41.

4. Ullmann N, Mirra V, Di Marco A, et al. Asthma: Differential Diagnosis and Comorbidities. Frontiers in Pediatrics 2018;6:276.

5. Scottish Intercollegiate Guidelines Network, British Thoracic Society. British guideline on the management of asthma: A national clinical guideline SIGN158 2019.

6. Larsen GL. Differences between adult and childhood asthma. Journal of Allergy and Clinical Immunology 2000;106:S153-S7.

7. Trivedi M, Denton E. Asthma in Children and Adults—What Are the Differences and What Can They Tell us About Asthma? Frontiers in Pediatrics 2019;7:256.

8. Bacharier LB, Boner A, Carlsen KH, et al. Diagnosis and treatment of asthma in childhood: a PRACTALL consensus report. Allergy 2008;63:5-34.

9. Asher I, Pearce N. Global burden of asthma among children. The International Journal of Tuberculosis and Lung Disease 2014;18:1269-78.

10. The Global Asthma Network. The Global Asthma Report. Auckland, New Zealand2018.

11. Asthma facts and statistics. 2021. (Accessed September, 2021, at https://www.asthma.org.uk/about/media/facts-and-statistics/.)

12. Lai CKW, Beasley R, Crane J, et al. Global variation in the prevalence and severity of asthma symptoms: Phase Three of the International Study of Asthma and Allergies in Childhood (ISAAC). Thorax 2009;64:476-83.

13. Licari A, Castagnoli R, Brambilla I, et al. Asthma Endotyping and Biomarkers in Childhood Asthma. Pediatr Allergy Immunol Pulmonol 2018;31:44-55.

14. Chung KF, Wenzel SE, Brozek JL, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. European Respiratory Journal 2014;43:343-73.

15. Stevens CA, Turner D, Kuehni CE, Couriel JM, Silverman M. The economic impact of preschool asthma and wheeze. European Respiratory Journal 2003;21:1000-6.

16. Bahadori K, Doyle-Waters MM, Marra C, et al. Economic burden of asthma: a systematic review. BMC Pulmonary Medicine 2009;9.

17. Battula M, Arunashekar P, Nagarajan VP. A Prospective Study to Assess the Quality of Life in Children with Newly Diagnosed Asthma and Their Caregivers using the Pediatric Asthma Quality of Life Questionnaire. Journal of Primary Care & Community Health 2020;11.

18. Expert Panel Report 3: Guidelines for the Diagnosis and Management of Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma–Summary Report 2007. Journal of Allergy and Clinical Immunology 2007;120:S94-S138.

19. Mims JW. Asthma: definitions and pathophysiology. International Forum of Allergy & Rhinology 2015;5:S2-S6.

20. Kudo M, Ishigatsubo Y, Aoki I. Pathology of asthma. Frontiers in Microbiology 2013;4:263.

21. Lloyd CM, Hessel EM. Functions of T cells in asthma: more than just T(H)2 cells. Nat Rev Immunol 2010;10:838-48.

22. Barnes PJ. Th2 cytokines and asthma: an introduction. Respiratory research 2001;2:64–5.

23. Kim ES, Kim SH, Kim KW, et al. Basement membrane thickening and clinical features of children with asthma. Allergy 2007;62:635-40.

24. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nature Medicine 2012;18:716-25.

25. Wenzel SE. Complex phenotypes in asthma: current definitions. Pulmonary Pharmacology & Therapeutics 2013;26:710-5.

26. Bantz SK, Zhu Z, Zheng T. The Atopic March: Progression from Atopic Dermatitis to Allergic Rhinitis and Asthma. Journal of Clinical & Cellular Immunology 2014;5.

27. Gold MS, Kemp AS. Atopic disease in childhood. Medical Journal of Australia 2005;182:298-304.

28. Romanet-Manent S, Charpin D, Magnan A, Lanteaume A, Vervloet D, Group EC. Allergic vs nonallergic asthma: what makes the difference? Allergy 2002;57:607-13.

29. Humbert M, Menz G, Ying S, et al. The immunopathology of extrinsic (atopic) and intrinsic (non-atopic) asthma: more similarities than differences. Immunology Today 1999;20:528-33.

30. Fitzpatrick AM, Teague WG, Meyers DA, et al. Heterogeneity of severe asthma in childhood: Confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. Journal of Allergy and Clinical Immunology 2011;127:382-9 e1-13.

31. Just J, Saint-Pierre P, Gouvis-Echraghi R, et al. Childhood Allergic Asthma Is Not a Single Phenotype. The Journal of Pediatrics 2014;164:815-20.

32. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. European Respiratory Journal 2012;40:55-60.

33. Howrylak JA, Fuhlbrigge AL, Strunk RC, Zeiger RS, Weiss ST, Raby BA. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. Journal of Allergy and Clinical Immunology 2014;133:1289-300.e12.

34. Oksel C, Granell R, Haider S, et al. Distinguishing Wheezing Phenotypes from Infancy to Adolescence. A Pooled Analysis of Five Birth Cohorts. Annals of the American Thoracic Society 2019;16:868-76.

35. Depner M, Fuchs O, Genuneit J, et al. Clinical and Epidemiologic Phenotypes of Childhood Asthma. American Journal of Respiratory and Critical Care Medicine 2013;189:129-38.

36. Lötvall J, Akdis CA, Bacharier LB, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. Journal of Allergy and Clinical Immunology 2011;127:355-60.

37. Kuruvilla ME, Lee FE, Lee GB. Understanding Asthma Phenotypes, Endotypes, and Mechanisms of Disease. Clinical Reviews in Allergy & Immunology 2019;56:219-33.

38. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and Wheezing in the First Six Years of Life. New England Journal of Medicine 1995;332:133-8.

39. 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;130:325-31.

40. Bacharier LB, Guilbert TW. Diagnosis and management of early asthma in preschool-aged children. Journal of Allergy and Clinical Immunology 2012;130:287-96; quiz 97-8.

41. Beasley R, Semprini A, Mitchell EA. Risk factors for asthma: is prevention possible? The Lancet 2015;386:1075-85.

42. Salam MT, Li Y-F, Langholz B, Gilliland FD. Early-life environmental risk factors for asthma: findings from the Children's Health Study. Environmental Health Perspectives 2004;112:760-5.

43. Bracken MB, Belanger K, Cookson WO, Triche E, Christiani DC, Leaderer BP. Genetic and Perinatal Risk Factors for Asthma Onset and Severity: A Review and Theoretical Analysis. Epidemiologic Reviews 2002;24:176-89.

44. London SJ, James Gauderman W, Avol E, Rappaport EB, Peters JM. Family history and the risk of early-onset persistent, early-onset transient, and late-onset asthma. Epidemiology 2001;12:577-83.

45. Burke W, Fesinmeyer M, Reed K, Hampson L, Carlsten C. Family history as a predictor of asthma risk. American Journal of Preventive Medicine 2003;24:160-9.

46. Litonjua AA, Carey VJ, Burge HA, Weiss ST, Gold DR. Parental History and the Risk for Childhood Asthma. American Journal of Respiratory and Critical Care Medicine 1998;158:176-81.

47. Arshad SH, Karmaus W, Raza A, et al. The effect of parental allergy on childhood allergic diseases depends on the sex of the child. Journal of Allergy and Clinical Immunology 2012;130:427-34 e6.

48. Ball TM, Castro-Rodriguez JA, Griffith KA, Holberg CJ, Martinez FD, Wright AL. Siblings, Day-Care Attendance, and the Risk of Asthma and Wheezing during Childhood. New England Journal of Medicine 2000;343:538-43.

49. Demenais F, Margaritte-Jeannin P, Barnes KC, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nature Genetics 2018;50:42-53.

50. Buniello A, MacArthur JA L, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research 2019;47:D1005-D12.

51. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. New England Journal of Medicine 2010;363:1211-21.

52. Pividori M, Schoettler N, Nicolae DL, Ober C, Im HK. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. The Lancet Respiratory Medicine 2019;7:509-22.

53. Belsky DW, Sears MR, Hancox RJ, et al. Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. The Lancet Respiratory Medicine 2013;1:453-61.

54. Bégin P, Nadeau KC. Epigenetic regulation of asthma and allergic disease. Allergy, Asthma & Clinical Immunology 2014;10:27.

55. Woodruff PG, Modrek B, Choy DF, et al. T-helper type 2-driven inflammation defines major subphenotypes of asthma. American Journal of Respiratory and Critical Care Medicine 2009;180:388-95.

56. Somineni HK, Zhang X, Biagini Myers JM, et al. Ten-eleven translocation 1 (TET1) methylation is associated with childhood asthma and traffic-related air pollution. Journal of Allergy and Clinical Immunology 2016;137:797-805 e5.

57. Hartwig FP, Loret de Mola C, Davies NM, Victora CG, Relton CL. Breastfeeding effects on DNA methylation in the offspring: A systematic literature review. PLoS One 2017;12.

58. Lockett GA, Soto-Ramirez N, Ray MA, et al. Association of season of birth with DNA methylation and allergic disease. Allergy 2016;71:1314-24.

59. Yang IV, Pedersen BS, Liu A, et al. DNA methylation and childhood asthma in the inner city. Journal of Allergy and Clinical Immunology 2015;136:69-80.

60. Joubert BR, Felix JF, Yousefi P, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. American Journal of Human Genetics 2016;98:680-96.

61. Dijk FN, Folkersma C, Gruzieva O, et al. Genetic risk scores do not improve asthma prediction in childhood. Journal of Allergy and Clinical Immunology 2019;144:857-60 e7.

62. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.

63. Meyers DA, Bleecker ER, Holloway JW, Holgate ST. Asthma genetics and personalised medicine. The Lancet Respiratory Medicine 2014;2:405-15.

64. Igartua C, Myers RA, Mathias RA, et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. Nature Communications 2015;6.

65. Morkve Knudsen T, Rezwan FI, Jiang Y, Karmaus W, Svanes C, Holloway JW. Transgenerational and intergenerational epigenetic inheritance in allergic diseases. Journal of Allergy and Clinical Immunology 2018;142:765-72.

66. Schatz M, Clark S, Camargo CA. Sex Differences in the Presentation and Course of Asthma Hospitalizations. Chest 2006;129:50-5.

67. Vink NM, Postma DS, Schouten JP, Rosmalen JG, Boezen HM. Gender differences in asthma development and remission during transition through puberty: the TRacking Adolescents'

Individual Lives Survey (TRAILS) study. Journal of Allergy and Clinical Immunology 2010;126:498-504 e1-6.

68. Forno E, Celedón JC. Health Disparities in Asthma. American Journal of Respiratory and Critical Care Medicine 2012;185:1033-5.

69. Jaakkola JJK, Gissler M. Maternal Smoking in Pregnancy, Fetal Development, and Childhood Asthma. American Journal of Public Health 2004;94:136-40.

70. Weitzman M, Gortmaker S, Walker DK, Sobol A. Maternal Smoking and Childhood Asthma. Pediatrics 1990;85:505.

71. Karmaus W, Dobai AL, Ogbuanu I, Arshard SH, Matthews S, Ewart S. Long-term effects of breastfeeding, maternal smoking during pregnancy, and recurrent lower respiratory tract infections on asthma in children. Journal of Asthma 2008;45:688-95.

72. Breton CV, Byun HM, Wenten M, Pan F, Yang A, Gilliland FD. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. American Journal of Respiratory and Critical Care Medicine 2009;180:462-7.

73. Custovic A, Custovic D, Kljaić Bukvić B, Fontanella S, Haider S. Atopic phenotypes and their implication in the atopic march. Expert Review of Clinical Immunology 2020;16:873-81.

74. Arshad SH, Tariq SM, Matthews S, Hakim E. Sensitization to Common Allergens and Its Association With Allergic Disorders at Age 4 Years: A Whole Population Birth Cohort Study. Pediatrics 2001;108:e33.

75. Arshad SH. Primary prevention of asthma and allergy. Journal of Allergy and Clinical Immunology 2005;116:3-14.

76. Du Toit G, Roberts G, Sayre PH, et al. Randomized trial of peanut consumption in infants at risk for peanut allergy. New England Journal of Medicine 2015;372:803-13.

77. Perkin MR, Logan K, Tseng A, et al. Randomized Trial of Introduction of Allergenic Foods in Breast-Fed Infants. New England Journal of Medicine 2016;374:1733-43.

78. Gdalevich M, Mimouni D, Mimouni M. Breast-feeding and the risk of bronchial asthma in childhood: A systematic review with meta-analysis of prospective studies. Journal of Pediatrics 2001;139:261-6.

79. Bloch AM, Mimouni D, Mimouni M, Gdalevich M. Does breastfeeding protect against allergic rhinitis during childhood? A meta-analysis of prospective studies. Acta Paediatrica 2002;91:275-9.

80. Friedman NJ, Zeiger RS. The role of breast-feeding in the development of allergies and asthma. Journal of Allergy and Clinical Immunology 2005;115:1238-48.

81. Duijts L. Fetal and infant origins of asthma. European Journal of Epidemiology 2012;27:5-14.

82. Oddy WH. Breastfeeding, Childhood Asthma, and Allergic Disease. Annals of Nutrition & Metabolism 2017;70 Suppl 2:26-36.

83. Brooks C, Pearce N, Douwes J. The hygiene hypothesis in allergy and asthma: an update. Current Opinion in Allergy and Clinical Immunology 2013;13:70-7.

84. Okada H, Kuhn C, Feillet H, Bach JF. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clin Exp Immunol 2010;160:1-9.

85. Bager P, Melbye M, Rostgaard K, Stabell Benn C, Westergaard T. Mode of delivery and risk of allergic rhinitis and asthma. Journal of Allergy and Clinical Immunology 2003;111:51-6.

86. Marfortt DA, Josviack D, Lozano A, Cuestas E, Aguero L, Castro-Rodriguez JA. Differences between preschoolers with asthma and allergies in urban and rural environments. Journal of Asthma 2018;55:470-6.

87. Kusel MM, de Klerk NH, Kebadze T, et al. Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma. Journal of Allergy and Clinical Immunology 2007;119:1105-10.

88. Jackson DJ, Gangnon RE, Evans MD, et al. Wheezing rhinovirus illnesses in early life predict asthma development in high-risk children. American Journal of Respiratory and Critical Care Medicine 2008;178:667-72.

89. Rzehak P, Wijga AH, Keil T, et al. Body mass index trajectory classes and incident asthma in childhood: Results from 8 European Birth Cohorts; a Global Allergy and Asthma European Network initiative. Journal of Allergy and Clinical Immunology 2013;131:1528-36.e13.

90. Ziyab AH, Karmaus W, Kurukulaaratchy RJ, Zhang H, Arshad SH. Developmental trajectories of Body Mass Index from infancy to 18 years of age: prenatal determinants and health consequences. Journal of Epidemiology and Community Health 2014;68:934-41.

91. Metsälä J, Kilkkinen A, Kaila M, et al. Perinatal factors and the risk of asthma in childhood--a population-based register study in Finland. American Journal of Epidemiology 2008;168:170-8.

92. Raby BA, Celedon JC, Litonjua AA, et al. Low-normal gestational age as a predictor of asthma at 6 years of age. Pediatrics 2004;114:e327-32.

93. Oh SM, Stefani KM, Kim HC. Development and application of chronic disease risk prediction models. Yonsei Medical Journal 2014;55:853-60.

94. Lutz CS, Huynh MP, Schroeder M, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health 2019;19:1659.

95. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One 2020;15.

96. Kucharski AJ, Russell TW, Diamond C, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases 2020;20:553-8.

97. Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683-90.

98. Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. Journal of Thrombosis and Haemostasis 2013;11 Suppl 1:129-41.

99. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011;343:d7163.

100. Vogenberg FR. Predictive and prognostic models: implications for healthcare decision-making in a modern recession. American Health & Drug Benefits 2009;2:218-22.

101. Lloyd-Jones DM. Cardiovascular risk prediction: basic concepts, current status, and future directions. Circulation 2010;121:1768-77.

102. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416.

103. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837-47.

104. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011;306:1688–98.

105. Goto T, Camargo CA, Jr., Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. American Journal of Emergency Medicine 2018;36:1650-4.

106. Adibi A, Sin DD, Safari A, et al. The Acute COPD Exacerbation Prediction Tool (ACCEPT): a modelling study. The Lancet Respiratory Medicine 2020;8:1013-21.

107. Burroughs AK, Sabin CA, Rolles K, et al. 3-month and 12-month mortality after first liver transplant in adults in Europe: predictive models for outcome. The Lancet 2006;367:225-32.

108. Smoots DW, Geyer JR, Lieberman DM, Berger MS. Predicting disease progression in childhood cerebellar astrocytoma. Child's Nervous System 1998;14:636-48.

109. Kothalawala DM, Kadalayil L, Weiss VBN, et al. Prediction models for childhood asthma: A systematic review. Pediatric Allergy and Immunology 2020;31:616-27.

110. Butler EM, Derraik JGB, Taylor RW, Cutfield WS. Prediction Models for Early Childhood Obesity: Applicability and Existing Issues. Hormone Research in Paediatrics 2018;90:358-67.

111. Yunginger JW, Reed CE, O'Connell EJ, Melton LJ, 3rd, O'Fallon WM, Silverstein MD. A community-based study of the epidemiology of asthma. Incidence rates, 1964-1983. American Review of Respiratory Disease 1992;146:888-94.

112. Annesi-Maesano I, Sterlin C, Caillaud D, et al. Factors related to under-diagnosis and under-treatment of childhood asthma in metropolitan France. Multidisciplinary Respiratory Medicine 2012;7.

113. Bush A, Fleming L, Saglani S. Severe asthma in children. Respirology 2017;22:886-97.

114. Loke YK, Blanco P, Thavarajah M, Wilson AM. Impact of Inhaled Corticosteroids on Growth in Children with Asthma: Systematic Review and Meta-Analysis. PLoS One 2015;10:e0133428.

115. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

116. Bui DS, Lodge CJ, Burgess JA, et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. The Lancet Respiratory Medicine 2018;6:535-44.

117. Waljee AK, Higgins PD, Singal AG. A primer on predictive models. Clinical and Translational Gastroenterology 2014;5:e44.

118. James G, Witten D, Hastie T, Tibshiran R. An Introduction to Statistical Learning. 1 ed: Springer-Verlag New York; 2013.

119. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Springer, New York, NY; 2009.

120. Alonso-Betanzos A, Bolon-Canedo V. Big-Data Analysis, Cluster Analysis, and Machine-Learning Approaches. Advances in Experimental Medicine and Biology 2018;1065:607-26.

121. Artetxe A, Beristain A, Grana M. Predictive models for hospital readmission risk: A systematic review of methods. Computer Methods and Programs in Biomedicine 2018;164:49-64.

122. Kloppel S, Stonnington CM, Chu C, et al. Automatic classification of MR scans in Alzheimer's disease. Brain 2008;131:681-9.

123. Singal AG, Mukherjee A, Elmunzer BJ, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. American Journal of Gastroenterology 2013;108:1723-30.

124. Kotsiantis SB. Supervised Machine Learning: A Review of Classification Techniques. Informatica 2007;31:249-68.

125. Dietterich T. Overfitting and undercomputing in machine learning. ACM Computing Surveys 1995;27:326-7.

126. Doroudi S. The Bias-Variance Tradeoff: How Data Science Can Inform Educational Debates. AERA Open 2020;6.

127. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825−30.

128. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biology 2019;20:76.

129. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. An International Journal on Information Fusion 2019;50:71-91.

130. Keogh E, Mueen A. Curse of Dimensionality. Encyclopedia of Machine Learning. Boston, MA: Springer US; 2011:257-8.

131. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70-9.

132. Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal 2018;19:179-89.

133. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;97:273-324.

134. He H, Garcia EA. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 2009;21:1263-84.

135. Gosain A, Sardana S. Handling Class Imbalance Problem using Oversampling Techniques: A Review. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)2017:79-85.

136. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2002;16:321–57.

137. Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing; 2005. p. 878-87.

138. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)2008:1322-8.

139. Elkan C. The Foundations of Cost-Sensitive Learning. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence; 2001. p. 973-78.

140. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.

141. Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus 2013;2:222.

142. Bennett DA. How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health 2001;25:464-9.

143. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 2013;3:e002847.

144. Batista GEAPA, Monard MC. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. Applied Artificial Intelligence 2002;17:519-33.

145. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research 2011;20:40-9.

146. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 2011;30:377-99.

147. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17:520–5.

148. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28:112-8.

149. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology 2019;110:63-73.

150. Bocchi L, Coppini G, Nori J, Valli G. Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. Medical Engineering & Physics 2004;26:303-12.

151. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 2010;10:16.

152. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics 2014;133:e54-63.

153. Hyland SL, Faltys M, Huser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine 2020;26:364-73.

154. Son YJ, Kim HG, Kim EH, Choi S, Lee SK. Application of support vector machine for prediction of medication adherence in heart failure patients. Healthcare Informatics Research 2010;16:253-9.

155. Patel SJ, Chamberlain DB, Chamberlain JM. A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage. Academic Emergency Medicine 2018;25:1463-70.

156. Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience 2017;9:329.

157. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. Scientific Reports 2017;7:2427.

158. Prosperi MC, Marinho S, Simpson A, Custovic A, Buchan IE. Predicting phenotypes of asthma and eczema with machine learning. BMC Medical Genomics 2014;7:S7.

159. Finkelstein J, Jeong IC. Machine learning approaches to personalize early prediction of asthma exacerbations. Annals of the New York Academy of Sciences 2017;1387:153-65.

160. Hinks TSC, Brown T, Lau LCK, et al. Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3–like protein 1. Journal of Allergy and Clinical Immunology 2016;138:61-75.

161. Krautenbacher N, Flach N, Bock A, et al. A strategy for high-dimensional multivariable analysis classifies childhood asthma phenotypes from genetic, immunological, and environmental factors. Allergy 2019;74:1364-73.

162. Goto T, Camargo CA, Jr., Faridi MK, Freishtat RJ, Hasegawa K. Machine Learning-Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. JAMA Network Open 2019;2:e186937.

163. Chatzimichail EA, Rigas AG, Paraskakis EN. An Artificial intelligence technique for the prediction of persistent asthma in children. Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine2010:1-4.

164. Chatzimichail E, Paraskakis E, Rigas A. An Evolutionary Two-Objective Genetic Algorithm for Asthma Prediction. 2013 UKSim 15th International Conference on Computer Modelling and Simulation2013:90-4.

165. 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.

166. Chatzimichail E, Paraskakis E, Rigas A. Predicting Asthma Outcome Using Partial Least Square Regression and Artificial Neural Networks. Advances in Artificial Intelligence 2013;2013:1-7.

167. Ahmad MA, Teredesai A, Eckert C. Interpretable machine learning in healthcare. 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018. p. 447.

168. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. p. 4768-77.

169. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135-44.

170. Arshad SH, Holloway JW, Karmaus W, et al. Cohort Profile: The Isle Of Wight Whole Population Birth Cohort (IOWBC). International Journal of Epidemiology 2018;47:1043-4i.

171. Asher MI, Keil U, Anderson HR, et al. International study of asthma and allergies in childhood (ISAAC): rationale and methods. European Respiratory Journal 1995;8:483-91.

172. Kurukulaaratchy RJ, Matthews S, Waterhouse L, Arshad SH. Factors influencing symptom expression in children with bronchial hyperresponsiveness at 10 years of age. Journal of Allergy and Clinical Immunology 2003;112:311-6.

173. Cole TJ. The LMS method for constructing normalized growth standards. European Journal of Clinical Nutrition 1990;44:45-60.

174. Ogbuanu IU, Karmaus W, Arshad SH, Kurukulaaratchy RJ, Ewart S. Effect of breastfeeding duration on lung function at age 10 years: a prospective birth cohort study. Thorax 2008;64:62-6.

175. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 2016;48:1279-83.

176. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics 2016;48:1443-8.

177. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics 2014;30:1266-72.

178. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature 2010;467:52-8.

179. Ziyab AH, Karmaus W, Yousefi M, et al. Interplay of filaggrin loss-of-function variants, allergic sensitization, and eczema in a longitudinal study covering infancy to 18 years of age. PLoS One 2012;7:e32721.

180. Beyan H, Down TA, Ramagopalan SV, et al. Guthrie card methylomics identifies temporally stable epialleles that are present at birth in humans. Genome Research 2012;22:2138-45.

181. Lehne B, Drong AW, Loh M, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biology 2015;16:37.

182. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30:1363-9.

183. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 2013;14:293.

184. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012;28:882-3.

185. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology 2016;17:208.

186. Custovic A, Simpson BM, Murray CS, Lowe L, Woodcock A. The National Asthma Campaign Manchester Asthma and Allergy Study. Pediatric Allergy and Immunology 2002;13:32-7.

187. Custovic A, Ainsworth J, Arshad H, 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;70:799-801.

188. Belgrave DCM, Simpson A, Semic-Jusufagic A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. Journal of Allergy and Clinical Immunology 2013;132:575-83 e12.

189. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research 2017;18:1-5.

190. Guyon I, Weston J, Barnhill S. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 2002;46:389-422.

191. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems 2006;83:83-90.

192. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genetics 2018;19:65.

193. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software 2010;36:1-13.

194. Kursa MB, Jankowski A, Rudnicki WR. Boruta – A System for Feature Selection. Fundamenta Informaticae 2010;101:271-85.

195. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning 1995;20:273-397.

196. Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Ratsch G. Support vector machines and kernels for computational biology. PLOS Computational Biology 2008;4:e1000173.

197. Kotsiantis SB. Decision trees: a recent overview. Artificial Intelligence Review 2011;39:261-83.

198. Breiman L. Random Forests. Machine Learning 2001;45:5–32.

199. Gardner MW, Dorling SR. Artificial neural networks (the multilayer perceptron) - A review of applications in the atmospheric sciences. Atmospheric Environment 1998;32:2627-36.

200. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 2012;13:281-305.

201. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011;45.

202. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.

203. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal 2014;35:1925-31.

204. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine 2013;4:627-35.

205. Steyerberg EW. Clinical Prediction Models. 2 ed2019.

206. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable 2019.

207. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

208. Purcell S, Chang C. PLINK 1.9.

209. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 2015;4.

210. Karim RB. Mixed Naive Bayes. 0.0.1 ed. Python Package Index2019.

211. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2020;2:56-67.

212. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering 2018;2:749-60.

213. Smit HA, Pinart M, Antó JM, et al. Childhood asthma prediction models: a systematic review. The Lancet Respiratory Medicine 2015;3:973-84.

214. Luo G, Nkoy FL, Stone BL, Schmick D, Johnson MD. A systematic review of predictive models for asthma development in children. BMC Medical Informatics and Decision Making 2015;15.

215. Colicino S, Munblit D, Minelli C, Custovic A, Cullinan P. Validation of childhood asthma predictive tools: A systematic review. Clinical & Experimental Allergy 2019;49:410-8.

216. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews 2015;4.

217. Reuters T. EndNote™ X8.2. 2017.

218. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Annals of Internal Medicine 2019;170:W1-W33.

219. Castro-Rodríguez JA, Holberg CJ, Wright AL, Martinez FD. A clinical index to define risk of asthma in young children with recurrent wheezing. American Journal of Respiratory and Critical Care Medicine 2000;162:1403-6.

220. Amin P, Levin L, Epstein T, et al. Optimum predictors of childhood asthma: persistent wheeze or the Asthma Predictive Index? Journal of Allergy and Clinical Immunology: In Practice 2014;2:709-15.

221. Singer F, Luchsinger I, Inci D, et al. Exhaled nitric oxide in symptomatic children at preschool age predicts later asthma. Allergy 2013;68:531-8.

222. Klaassen EMM, van de Kant KDG, Jöbsis Q, et al. Exhaled Biomarkers and Gene Expression at Preschool Age Improve Asthma Prediction at 6 Years of Age. American Journal of Respiratory and Critical Care Medicine 2015;191:201-7.

223. Chang TS, Lemanske RF, Guilbert TW, et al. Evaluation of the Modified Asthma Predictive Index in High-Risk Preschool Children. The Journal of Allergy and Clinical Immunology: In Practice 2013;1:152-6.

224. Guilbert TW, Morgan WJ, Zeiger RS, et al. Atopic characteristics of children with recurrent wheezing at high risk for the development of childhood asthma. Journal of Allergy and Clinical Immunology 2004;114:1282-7.

225. Hafkamp-de Groen E, Lingsma HF, Caudri D, et al. Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. Journal of Allergy and Clinical Immunology 2013;132:1303-10.e6.

226. Lødrup Carlsen KC, Söderström L, Mowinckel P, et al. Asthma prediction in school children; the value of combined IgE-antibodies and obstructive airways disease severity score. Allergy 2010;65:1134-40.

227. 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. Pediatric Pulmonology 2011;46:1175-81.

228. Grabenhenrich LB, Reich A, Fischer F, et al. The novel 10-item asthma prediction tool: external validation in the German MAS birth cohort. PLoS One 2014;9:e115852.

229. Pedersen ESL, Spycher BD, de Jong C, et al. The Simple 10-Item Predicting Asthma Risk in Children Tool to Predict Childhood Asthma-An External Validation. The Journal of Allergy and Clinical Immunology: In Practice 2019;7:943-53.e4.

230. Leonardi NA, Spycher BD, Strippoli MP, 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;127:1466-72 e6.

231. Caudri D, Wijga A, A. Schipper CM, et al. Predicting the long-term prognosis of children with symptoms suggestive of asthma at preschool age. Journal of Allergy and Clinical Immunology 2009;124:903-10.e7.

232. 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. Journal of Asthma 2011;48:1015-21.

233. Pescatore AM, Dogaru CM, Duembgen L, et al. A simple asthma prediction tool for preschool children with wheeze or cough. Journal of Allergy and Clinical Immunology 2014;133:111-8.e13.

234. Biagini Myers JM, Schauberger E, He H, et al. A Pediatric Asthma Risk Score to better predict asthma development in young children. Journal of Allergy and Clinical Immunology 2018;143:1803-10.e2.

235. Kurukulaaratchy RJ, Matthews S, Holgate ST, Arshad SH. Predicting persistent disease among children who wheeze during early life. European Respiratory Journal 2003;22:767-71.

236. Eysink PE, ter Riet G, Aalberse RC, 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;55:125-31.

237. Devulapalli CS, Carlsen KCL, Haland G, et al. Severity of obstructive airways disease by age 2 years predicts asthma at 10 years of age. Thorax 2008;63:8-13.

238. Lodrup Carlsen KC, Mowinckel P, Granum B, Carlsen KH. Can childhood asthma be predicted at birth? Clinical & Experimental Allergy 2010;40:1767-75.

239. van der Mark LB, van Wonderen KE, Mohrs J, van Aalderen WMC, ter Riet G, Bindels PJ. Predicting asthma in preschool children at high risk presenting in primary care: development of a clinical asthma prediction score. Primary Care Respiratory Journal 2014;23:52-9.

240. 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;52:729-36.

241. Szentpetery SS, Gruzieva O, Forno E, et al. Combined effects of multiple risk factors on asthma in school-aged children. Respiratory Medicine 2017;133:16-21.

242. Wang R, Simpson A, Custovic A, Foden P, Belgrave D, Murray CS. Individual risk assessment tool for school-age asthma prediction in UK birth cohort. Clin Exp Immunol 2019;49:292-8.

243. Aligne CA, Auinger P, Byrd RS, Weitzman M. Risk Factors for Pediatric Asthma Contributions of Poverty, Race, and Urban Residence. American Journal of Respiratory and Critical Care Medicine 2000;162:873–7.

244. Elphick HE, Sherlock P, Foxall G, et al. Survey of respiratory sounds in infants. Archives of Disease in Childhood 2001;84:35-9.

245. Cane RS, McKenzie SA. Parents' interpretations of children's respiratory symptoms on video. Archives of Disease in Childhood 2001;84:31-4.

246. Sly PD, Boner AL, Björksten B, et al. Early identification of atopy in the prediction of persistent asthma in children. The Lancet 2008;372:1100-6.

247. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. Perspectives in Clinical Research 2017;8:148-51.

248. Wang H, Peng J, Wang B, et al. Inconsistency Between Univariate and Multiple Logistic Regressions. Shanghai Archives of Psychiatry 2017;29:124-8.

249. Van Wonderen KE, Van Der Mark LB, Mohrs J, Bindels PJ, Van Aalderen WM, Ter Riet G. Different definitions in childhood asthma: how dependable is the dependent variable? European Respiratory Journal 2010;36:48-56.

250. Patel D, Hall GL, Broadhurst D, Smith A, Schultz A, Foong RE. Does machine learning have a role in the prediction of asthma in children? Paediatric Respiratory Reviews 2021.

251. Owora AH, Tepper RS, Ramsey CD, Becker AB, Genuneit J. Decision tree-based rules outperform risk scores for childhood asthma prognosis. Pediatric Allergy and Immunology 2021.

252. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Medicine 2013;10:e1001381.

253. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. Journal of Clinical Epidemiology 2003;56:826-32.

254. Price DB, Rigazio A, Campbell JD, et al. Blood eosinophil count and prospective annual asthma disease burden: a UK cohort study. The Lancet Respiratory Medicine 2015;3:849-58.

255. Saglani S, Custovic A. Childhood Asthma: Advances Using Machine Learning and Mechanistic Studies. American Journal of Respiratory and Critical Care Medicine 2019;199:414-22.

256. Fehrenbach H, Smolinska A, Klaassen EMM, et al. Profiling of Volatile Organic Compounds in Exhaled Breath As a Strategy to Find Early Predictive Signatures of Asthma in Children. PLoS ONE 2014;9.

257. AlSaad R, Malluhi Q, Janahi I, Boughorbel S. Interpreting patient-Specific risk prediction using contextual decomposition of BiLSTMs: application to children with asthma. BMC Medical Informatics and Decision Making 2019;19:214.

258. Bose S, Kenyon CC, Masino AJ. Personalized prediction of early childhood asthma persistence: A machine learning approach. Plos One 2021;16.

259. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology 2014;25:114-21.

260. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007;8:25.

261. Kothalawala DM, Murray CS, Simpson A, et al. Development of childhood asthma prediction models using machine learning approaches. Clinical and Translational Allergy 2021;11.

262. Carr TF, Kraft M. Use of biomarkers to identify phenotypes and endotypes of severe asthma. Annals of Allergy, Asthma & Immunology 2018;121:414-20.

263. Stephenson L. Monoclonal Antibody Therapy for Asthma. Clinical Pulmonary Medicine 2017;24:250-7.

264. Ivanova O, Richards LB, Vijverberg SJ, et al. What did we learn from multiple omics studies in asthma? Allergy 2019;74:2129-45.

265. Ullemar V, Magnusson PK, Lundholm C, et al. Heritability and confirmation of genetic association studies for childhood asthma in twins. Allergy 2016;71:230-8.

266. Thomsen SF, Van Der Sluis S, Kyvik KO, Skytthe A, Backer V. Estimates of asthma heritability in a large twin sample. Clinical & Experimental Allergy 2010;40:1054-61.

267. Reese SE, Xu CJ, den Dekker HT, et al. Epigenome-wide meta-analysis of DNA methylation and childhood asthma. Journal of Allergy and Clinical Immunology 2019;143:2062-74.

268. Edris A, den Dekker HT, Melén E, Lahousse L. Epigenome-wide association studies in asthma: A systematic review. Clinical and Experimental Allergy 2019;49:953-68.

269. Qi C, Xu C-J, Koppelman GH. The role of epigenetics in the development of childhood asthma. Expert Review of Clinical Immunology 2019;15:1287-302.

270. El-Husseini ZW, Gosens R, Dekker F, Koppelman GH. The genetics of asthma and the promise of genomics-guided drug target discovery. The Lancet Respiratory Medicine 2020;8:1045-56.

271. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31:3555-7.

272. Ferreira MAR, Mathur R, Vonk JM, et al. Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. American Journal of Human Genetics 2019;104:665-84.

273. Myers TA, Chanock SJ, Machiela MJ. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Frontiers in Genetics 2020;11:157.

274. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nature Protocols 2020;15:2759-72.

275. Affinito O, Palumbo D, Fierro A, et al. Nucleotide distance influences co-methylation between nearby CpG sites. Genomics 2020;112:144-50.

276. Martin TC, Yet I, Tsai P-C, Bell JT. coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. BMC Bioinformatics 2015;16.

277. Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. Epigenetics 2019;15:1-11.

278. Fernández-Sanlés A, Sayols-Baixeras S, Curcio S, Subirana I, Marrugat J, Elosua R. DNA Methylation and Age-Independent Cardiovascular Risk, an Epigenome-Wide Approach. Arteriosclerosis, Thrombosis, and Vascular Biology 2018;38:645-52.

279. Guan Z, Raut JR, Weigl K, et al. Individual and joint performance of DNA methylation profiles, genetic risk score and environmental risk scores for predicting breast cancer risk. Molecular Oncology 2019;14:42-53.

280. Yu H, Raut JR, Schöttker B, Holleczek B, Zhang Y, Brenner H. Individual and joint contributions of genetic and methylation risk scores for enhancing lung cancer risk stratification: data from a population-based cohort in Germany. Clinical Epigenetics 2020;12.

281. Elliott HR, Tillin T, McArdle WL, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clinical Epigenetics 2014;6.

282. Chowdhury D, Zhou X, Li B, et al. Editorial: Predicting High-Risk Individuals for Common Diseases Using Multi-Omics and Epidemiological Data. Frontiers in Genetics 2021;12.

283. Odintsova VV, Rebattu V, Hagenbeek FA, et al. Predicting Complex Traits and Exposures From Polygenic Scores and Blood and Buccal DNA Methylation Profiles. Frontiers in Psychiatry 2021;12.

284. Fontanillas P, Alipanahi B, Furlotte NA, et al. Disease risk scores for skin cancers. Nature Communications 2021;12:160.

285. Clark H, Granell R, Curtin JA, et al. Differential associations of allergic disease genetic variants with developmental profiles of eczema, wheeze and rhinitis. Clinical and Experimental Allergy 2019;49:1475-86.

286. Park J, Jang H, Kim M, et al. Predicting allergic diseases in children using genome-wide association study (GWAS) data and family history. World Allergy Organization Journal 2021;14.

287. Simard M, Madore A-M, Girard S, et al. Polygenic risk score for atopic dermatitis in the Canadian population. Journal of Allergy and Clinical Immunology 2021;147:406-9.

288. Sordillo JE, Lutz SM, McGeachie MJ, et al. Pharmacogenetic Polygenic Risk Score for Bronchodilator Response in Children and Adolescents with Asthma: Proof-of-Concept. Journal of Personalized Medicine 2021;11.

289. Spycher BD, Henderson J, Granell R, et al. Genome-wide prediction of childhood asthma and related phenotypes in a longitudinal birth cohort. Journal of Allergy and Clinical Immunology 2012;130:503-9.e7.

290. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics 2018;19:371-84.

291. Jylhävä J, Pedersen NL, Hägg S. Biological Age Predictors. EBioMedicine 2017;21:29-36.

292. Sugden K, Hannon EJ, Arseneault L, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Translational Psychiatry 2019;9.

293. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. Epigenomics 2019;11:1469-86.

294. Onwuka JU, Li D, Liu Y, et al. A panel of DNA methylation signature from peripheral blood may predict colorectal cancer susceptibility. BMC Cancer 2020;20.

295. Clark SL, Hattab MW, Chan RF, et al. A methylation study of long-term depression risk. Molecular Psychiatry 2019;25:1334-43.

296. Hamilton OKL, Zhang Q, McRae AF, et al. An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. International Journal of Obesity 2019;43:1795-802.

297. Jiang Y, Wei J, Zhang H, et al. Epigenome wide comparison of DNA methylation profile between paired umbilical cord blood and neonatal blood on Guthrie cards. Epigenetics 2019;15:454-61.

298. Deng Y, Wan H, Tian J, et al. CpG-methylation-based risk score predicts progression in colorectal cancer. Epigenomics 2020;12:605-15.

299. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Medicine 2020;12.

300. Janssens ACJW, Joyner MJ. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? Clinical Chemistry 2019;65:609-11.

301. Machiela MJ, Chen C-Y, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genetic Epidemiology 2011;35:506-14.

302. Song S, Jiang W, Hou L, Zhao H. Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. PLOS Computational Biology 2020;16.

303. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. European Journal of Epidemiology 2018;33:459-64.

304. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. The Lancet Respiratory Medicine 2015;3:42-52.

305. Morales E, Duffy D. Genetics and Gene-Environment Interactions in Childhood and Adult Onset Asthma. Frontiers in Pediatrics 2019;7.

306. McCarthy M, Birney E. Personalized profiles for disease risk must capture all facets of health. Nature 2021;597:175-7.

307. Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. OMICS 2018;22:630-6.

308. Lin E, Lane H-Y. Machine learning and systems genomics approaches for multi-omics data. Biomarker Research 2017;5.