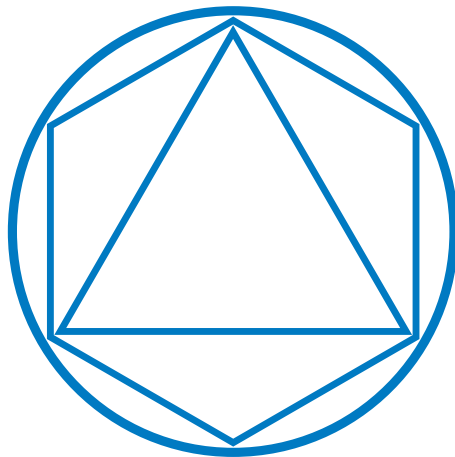

Learning on complex, biased, and big data:
disease risk prediction in epidemiological studies
and genomic medicine on the example of
childhood asthma



Norbert Krautenbacher

Juli 2018

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik — Lehrstuhl M12 (Mathematische Modellierung
biologischer Systeme)

**Learning on complex, biased, and big data: disease risk prediction
in epidemiological studies and genomic medicine on the example of
childhood asthma**

Norbert Krautenbacher

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen
Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:

Prof. Dr. Christina Kuttler

Prüfer der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Anne-Laure Boulesteix,
Ludwig-Maximilians-Universität München
3. Prof. Dr. Barbara Hammer, Universität Bielefeld

Die Dissertation wurde am 19.07.2018 bei der Technischen Universität München
eingereicht und durch die Fakultät für Mathematik am 12.09.2018 angenommen.

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die zur Entstehung dieser Doktorarbeit beigetragen haben. Besonders möchte ich mich bedanken bei ...

... Fabian Theis, meinem Doktorvater, für das Ermöglichen einer wissenschaftlichen Arbeit auf einem interdisziplinären Gebiet zwischen Statistik, Mathematik und Biologie, die Förderung und das Vertrauen.

... Christiane Fuchs, meiner Betreuerin, für die uneingeschränkte Unterstützung bei sämtlichen Angelegenheiten, die stets angenehme Arbeitsatmosphäre, die fachlichen Diskussionen, den Feiraum und das Fördern meiner Ideen.

... meiner Prüfungskommission, den beiden Gutachtern Anne-Laure Boulesteix und Barbara Hammer sowie der Vorsitzenden Christina Kuttler.

... meinen Kollaborationspartnern, insbesondere Markus Ege, Bianca Schaub und Andreas Böck für die äußerst interessante Zusammenarbeit.

... Donna Ankerst für viele Tipps und den wertvollen fachlichen Input.

... Hannes Petermeier, der mir ein sehr angenehmes Lehren neben der Doktorarbeit ermöglicht hat, so dass genügend Freiraum zum Forschen blieb.

... Nicolai Flach für den tollen Einsatz bei einem meiner Projekte im Rahmen seiner Masterarbeit.

... meinen Kollegen und Freunden am ICB, insbesondere Atefeh, Adriana, Lisa B., Michael L., Michael S., Hans, Valerio, Lisa A., Hannah, Susanne und den restlichen Mitgliedern der Biostatistikgruppe sowie dem "A Bavarian Dream"-Team für eine sehr schöne Zeit und eine wunderbar entspannte Atmosphäre am Institut.

... meinen Freunden außerhalb des Instituts sowie meiner Familie für den wertvollen seelischen Beistand und besonders meinen Eltern auch für das Ermöglichen des Statistikstudiums, welches Voraussetzung für diese Promotion war.

Abstract

Predicting the risk of complex diseases is a field of growing relevance in medicine and shows high potential of refinement and improvement by integrating new data types and larger data sets. But it also creates new challenges.

In this thesis, we investigate and overcome issues on four of these challenges: (i) *complex missing data structures* where state-of-the-art methods are not applicable, (ii) *sample selection bias* occurring when data was not taken at random from a population, (iii) *high-dimensional data*, when numbers of variables are tens of thousands to millions and thus much higher than numbers of observations, and (iv) *multi-omics data* describing data consisting of several groups of variables which are of different types and of different dimensions and each of these groups should be integrated appropriately.

First, we look at (i) and especially at (ii) by investigating how to correct arbitrary statistical and machine learning models for sample selection bias resulting from a special study design. We develop two novel correction approaches and compare these with, partly extended, state-of-the-art correction methods in theory, in a simulation study, and on real data. One of our proposed approaches outperforms all others in the case of the random forest classifier and performs at least as well as other correction methods for other classifiers. We provide the implemented methods in terms of a publicly available Software.

Second, we address (ii), especially (iii), and (iv); data from a big European study which contain farm-related environmental and genome-wide genetic variables are analyzed with the goal of predicting the risk of childhood asthma. Modern statistical learning approaches are applied for handling high-dimensional data and incorporating multi-omics structure appropriately. For the learning procedure we take into account sample selection bias which is present in the data and propose methodology allowing for correcting the precision of estimates for performance and comparing performances of two classifiers correctly. We identify family history of childhood asthma as most predictive variable and by external validation find only moderate prediction power which, however, is higher when only farm-children are used for analyses.

Eventually, we deal with a project for (iii) and in particular the combination of (i) and (iv) where the goal is to distinguish between allergic asthmatics, non-allergic asthmatics and healthy controls by again using environmental and (a selection of) genetic variables but in addition diverse immunological variables. Here, seven types of data are present in the frame of a complex missing data structure. We employ statistical and machine learning classifiers for predicting the disease. Incorporating these in a novel modeling strategy allows to use all information in the data rather than leaving out certain values as classical naive strategies would do. Applying several strategies taking into account different scales

and granularities in the data yields good performance and precision of performance is increased when the proposed strategy is applied. We identify novel predictive variables from the immunological data sets, especially three genes which have not been associated with childhood asthma in literature before.

Altogether we show how prediction for diseases can be improved by utilizing statistical methodology for taking into account bias, complexity and bigness of data.

Zusammenfassung

Die Prädiktion des Risikos komplexer Krankheiten in der Medizin gewinnt an Relevanz und birgt hohes Potential der Verfeinerung und Verbesserung, indem neue Datentypen sowie größere Datensätze integriert werden. Ebenso bringt sie jedoch neue Herausforderungen mit sich.

In dieser Arbeit untersuchen und lösen wir Aspekte zu vier dieser Herausforderungen: (i) *komplexe Datenstrukturen aufgrund fehlender Werte*, bei welchen herkömmliche Methoden nicht anwendbar sind, (ii) *Sample-Selection-Bias* (Selektionsverzerrung), welcher auftritt, wenn Daten nicht zufällig von einer Grundgesamtheit gezogen wurden, (iii) *hochdimensionale Daten*, wenn die Anzahl von Variablen in der Größenordnung von Zehntausenden bis Millionen liegt und damit wesentlich größer als die Anzahl von Beobachtungen ist, sowie (iv) den Fall von *Multi-Omics-Daten*, bei welchem Daten aus Gruppen von Variablen unterschiedlicher Typen sowie unterschiedlicher Dimensionen bestehen und jede dieser Gruppen angemessen einbezogen werden soll.

Zunächst betrachten wir (i) und besonders (ii), indem wir untersuchen, wie beliebige statistische bzw. maschinelle Lernmethoden für Sample-Selection-Bias korrigiert werden können, wenn dieser durch ein spezielles Studiendesign hervorgerufen wurde. Wir entwickeln zwei neue Korrekturmethode und vergleichen diese mit, zum Teil erweiterten, klassischen Korrekturmethode in der Theorie, in einer Simulationsstudie und auf echten Daten. Eine unserer entwickelten Methoden übertrifft alle anderen im Falle des Random-Forest-Klassifikators und zeigt bei anderen Klassifikationsmodellen eine mindestens ebenso große Güte wie andere Korrekturmethode. Die implementierten Methoden stellen wir in Form von öffentlich zugänglicher Software zur Verfügung.

Im Weiteren adressieren wir (ii), besonders (iii), sowie (iv); Daten einer großen europäischen Studie, welche bauernhofbezogene Umwelt- sowie genomweite Genetikvariablen enthält, werden analysiert. Das Ziel ist es hierbei, das Risiko von Asthma im Kindesalter vorherzusagen. Hierzu werden moderne statistische Lernmethoden, welche hochdimensionale Daten handhaben und Multi-Omics-Strukturen einbinden können, eingesetzt. Innerhalb der Lernprozedur beziehen wir den in den Daten vorliegenden Sample-Selection-Bias ein und schlagen eine Methodik vor, welche es erlaubt, die Präzision von Schätzern für die Prädiktionsgüte zu korrigieren sowie das Verhalten zweier Klassifikatoren korrekt zu vergleichen. Wir identifizieren Familienanamnese von Asthma im Kindesalter als prädiktivste Variable und finden bei externer Validierung nur mäßige Vorhersagekraft, welche sich allerdings verbessert, wenn Analysen nur auf Bauernhofkindern durchgeführt werden.

Zu (iii) und insbesondere zu (i) und (iv) kombiniert behandeln wir schließlich ein Projekt,

dessen Ziel es ist, zwischen allergischen Asthmatikern, nicht-allergischen Asthmatikern und gesunden Kontrollen zu unterscheiden, erneut unter Nutzung von Umwelt- und (einer Auswahl von) Genetikvariablen, jedoch zusätzlich von immunologischen Variablen. Hierbei liegen sieben Datentypen im Rahmen einer komplexen Struktur fehlender Daten vor. Integriert in eine neuartige Modellierungsstrategie setzten wir Klassifikatoren aus dem statistischen bzw. maschinellen Lernen ein, um die Krankheit vorherzusagen. Diese erlaubt es, sämtliche Information in den Daten zu nutzen anstatt gewisse Datenwerte auszulassen, wie es bei naiveren Strategien der Fall wäre. Die Anwendung verschiedener Strategien, welche die unterschiedlichen Skalen und Granularitäten in den Daten einbeziehen, führt zu gutem Prädiktionsverhalten und einer erhöhten Präzision der Schätzer für die Prädiktionsgüte, wenn die vorgeschlagene Strategie eingesetzt wird. Wir identifizieren prädiktive Variablen aus den immunologischen Datensätzen, insbesondere drei Gene, welche in der Literatur bisher nicht mit Asthma im Kindheitsalter in Verbindung gebracht worden sind. Insgesamt zeigen wir, wie die Prädiktion von Krankheiten mithilfe des Einsatzes von statistischer Methodik, welche Verzerrung, Komplexität und Größe der Daten berücksichtigt, verbessert werden kann.

Contents

1	Introduction	1
1.1	Predicting disease risk	1
1.2	Childhood asthma	3
1.3	Challenges in disease risk estimation	5
1.4	Objectives of the thesis	7
1.5	Contributing manuscripts	8
2	Methodological background	11
2.1	General notation	11
2.2	Imputation of missing data	12
2.2.1	Types of missing data	12
2.2.2	Single imputation techniques	13
2.2.3	Multiple imputation	14
2.3	Statistical and machine learning models	17
2.3.1	Logistic regression	18
2.3.2	Multinomial regression	18
2.3.3	Penalized regression	19

2.3.4	Integrative L_1 -penalized regression with penalty factors based on multi-omics data	20
2.3.5	Classification trees	21
2.3.6	Random forest	22
2.3.7	Gradient boosting	23
2.3.8	Naive Bayes	26
2.4	Validating prediction models	26
2.4.1	Measuring performance	26
2.4.2	Independent validation	28
3	Correcting classifiers for sample selection bias in two-phase case-control studies	31
3.1	Sample selection bias and stratified random sampling	35
3.1.1	Sample selection bias — definition	35
3.1.2	Two-phase case-control studies	36
3.1.3	Stratified random samples	37
3.2	Methods	38
3.2.1	State of the art correction approaches	39
3.2.2	Correcting covariance structures	41
3.2.3	Properties of correction approaches	45
3.2.4	Classifiers	47
3.3	Simulation study	48
3.3.1	Design	48
3.3.2	Results	52
3.4	Real data application	56

<i>CONTENTS</i>	xiii
3.4.1 Design	59
3.4.2 Results	60
3.5 Discussion and Conclusion	62
3.6 Additional Material	65
4 Encountering big data: predicting childhood asthma risk by genetic and environmental variables	67
4.1 Data collection	68
4.1.1 Population and questionnaires	68
4.1.2 Genotyping	70
4.2 Computational and statistical analysis	71
4.3 Correcting and comparing losses for sample bias	75
4.4 Results	76
4.5 Discussion	83
5 Tackling multi-omics missingness patterns: classifying childhood asthma phenotypes using genetics, environment and immunology	93
5.1 Data collection	95
5.1.1 Study population	95
5.1.2 Modalities	95
5.1.3 Genotyping	95
5.1.4 Microarrays	96
5.1.5 RT-qPCR, flow cytometry and cytokines	96
5.2 Computational and statistical analysis	96
5.3 A strategy for prediction on incomplete multi-omics data	102

5.4	Results	102
5.4.1	Prediction modeling	102
5.4.2	Variable importance	105
5.5	Discussion	107
6	Summary and perspectives	115
6.1	Summary	115
6.2	Outlook	117
6.3	Conclusions	121
A	Supplementary Material	151
A.1	On correcting classifiers for sample selection bias	151
A.1.1	Simulation scenarios with variables from different distribution families	151
A.1.2	Investigation of varying sample size and degree of imbalance	152
A.1.3	Investigation of other bias types	155
A.2	On analyses on GABRIELA study	157
A.2.1	P-value adjustment for variable importance	157
A.2.2	Variable importance genome-wide	161
A.2.3	Tables on family history, environment and genetics	161
A.2.4	External validation for non-farm and farm children	164
A.2.5	Parametric inverse-probability bagging on GABRIELA	166
A.3	On analyses on CLARA study	167
A.3.1	Wider selection of important variables	167

Chapter 1

Introduction

1.1 Predicting disease risk

Predicting complex diseases is an highly topical field in biomedicine. In medical practice one hopes to improve disease risk management, that is diagnosis, treatment, and prevention, on an individual level by using knowledge on all kinds of risk factors. By this, health care on an individual level can be improved [73]. This goal is accompanied with the concept of *precision medicine* which is “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person” [11], thus a disease can be defined at a higher resolution, so that subgroups of diseases can be targeted more accurately with new therapies [10]. As Figure 1.1 illustrates the approach of precision medicine overcomes the traditional approach of assigning the appropriate treatment for an individual. Predicting or classifying disease risks can be seen as a first component of precision medicine (see Figure 1.1). Therefore, various types of measurable possible influence factors are taken into account with the aim to build powerful statistical models to predict whether or with which probability an individual will suffer from a disease. Such factors can be demographical variables like age or place of domicile. Environmental exposure is a further important category of influential factors for many diseases [133, 156]. An often very predictive factor is the information on family members having suffered from the disease [164]. As proven in literature [41], family history may partly cover but surely not replace the predictiveness of another non-neglectable factor which becomes more and more useful for risk prediction: genetics. Especially since the sequence of the human genome has been completed in the beginning of the 21st century — the sequence of the last chromosome was published in 2006 [68] — prediction of disease


























	<i>Traditional Approach</i>	<i>Precision Medicine Approach</i>				
Population of Individuals						
a) Classify by Risk						
b) Surveillance for Preclinical Disease						
Signs or Symptoms						
c) Treat with						
Strategy	"One Size Fits All" Leads to Overall Mixed Results			Focus Existing	Repurpose FDA Approval	Invent New
						
Outcome						
	Benefit	No Effect	Adverse	Benefit	Benefit	Benefit

Figure 1.1: Scheme of the traditional medicine approach compared to precision medicine. The latter consists of three key elements: a) classifying by risk which is the focus of this thesis, b) surveillance for preclinical disease, and c) aligning an extended repertoire of treatments with the individual's molecular drivers of disease (adapted from Asher et al. [9]).

risk or generally of complex phenotypic traits by genetics up to genome wide prediction modeling has become a meaningful subject of biomedical research. Successful prediction models by using data on single-nucleotide polymorphisms (SNPs), which are variations in single nucleotides occurring at certain positions in the genome, have been built on type 1 diabetes, for instance [60, 184, 185].

In order to make an estimation of the risk of an individual, i.e. for which a future disease status should be predicted, cohort study data is used where for each individual variables are measured and the disease status is known. Classical statistical models and modern machine learning models are then used to learn on this given data, that is these models are fitted on the given data and can be applied on new data where the outcome has to be predicted.

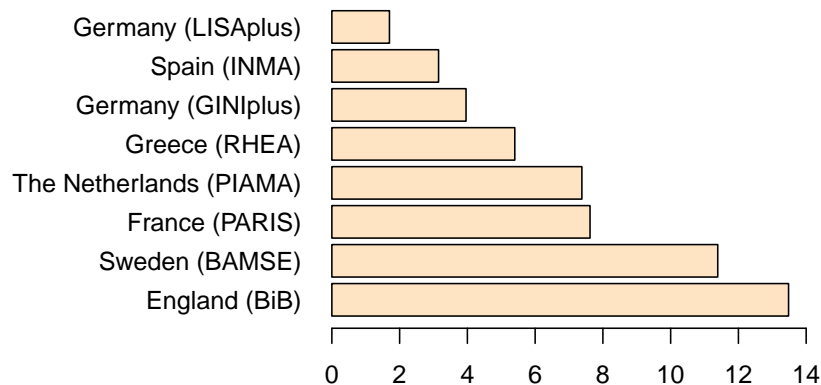


Figure 1.2: (Unadjusted) prevalence rates (percentages) of ever asthma at age 4 years in Europe by cohort. Values extracted from Uphoff et al. [172].

1.2 Childhood asthma

In this thesis childhood asthma will play a special role, as we will mostly focus our investigation of disease risk prediction on this case — within a collaboration with two groups of clinicians at the Dr. von Hauner Children’s Hospital in Munich.

Asthma in the general context can be seen as a collection of diseases rather than one specific disease which is a first point why prediction of asthma is difficult [40]. Therefore in the applications of the thesis we will define asthma phenotypes in several ways (Chapter 4 and 5). According to Uphoff et al. [172], based on MeDALL cohorts in Europe, 4 year old aged asthma prevalence ranges from 1.72% in a German study (LISApplus) up to 13.48% in England (see also Figure 1.2 for prevalence rates in Europe). However, as childhood asthma is not unambiguously defined and thus not diagnosed consistently the prevalence of wheeze is to be mentioned as well: the proportion of children who ever had wheezing or whistling in the chest varies from 9.82% in Greece to 55.37% in Spain [172]. These varying numbers may suggest that childhood asthma may not be diagnosed the same way in different countries — for instance, a study has shown that asthmatic complaints led to a diagnosis for asthma for Dutch children, but to a diagnosis of bronchitis in German children [124]. Nevertheless, independent of that childhood asthma can be seen as generally very common disease.

For this disease many possible predictors have been investigated. Childhood asthma is highly heritable [171] which makes family history a predictive variable. Heritability gives information on “how much of the variation in a trait is due to variation in genetic factors” [188], or is formally defined “as the proportion of phenotypic variation [...] that is due to variation in genetic values [...]” [188]. According to Ullemer et al. [171] childhood asthma has a heritability of 0.82 (95% CI 0.79–0.85) which means that 82% of the childhood asthma phenotype variability can be attributed to variation in genes.

Further it has been shown that demographical factors, such as sex [66], play a role in the risk of getting asthma. According to this, boys generally have a higher risk of getting childhood asthma than girls.

Several publications give evidence that childhood asthma is partly caused by environmental exposure, especially farming environment plays a role and protects from childhood asthma [49, 63].

Belsky et al. [15], Ege et al. [49], Moffatt et al. [123], Ober and Hoffjan [130], Vercelli [178] investigate the genetic influence and Moffatt et al. [123] concludes asthma as being a genetically heterogeneous disease and that determination of individual risk of asthma is difficult.

Ege and Strachan [48], Ege et al. [49], for instance, have performed a gene-environment interaction analysis; in general, the analysis of these kind of interactions is a complex issue (and one central goal of Helmholtz Zentrum München). Some of the difficulties are that data have to be large to guarantee sufficient power for a high number of tests of significance, measuring environmental exposure is generally complex, in particular the incorporation of its temporality, and that genetic variation is limited [116]. Ege and Strachan [48], Ege et al. [49] have found associations with this kind of predictors in their interaction analyses; strong effects have been found for the two-way interactions of several genes with farming exposure, the consumption of raw farm milk, or contact with cows and straw [49].

Further, Raedler et al. [139] detect immunological factors that influence the development of childhood asthma.

These findings indicate that childhood asthma represents a disease caused by several types of predictors. Thus, studies on childhood asthma represent an ideal application of risk prediction analysis comprising all the challenges this thesis aims to investigate and to solve as described as follows.

1.3 Challenges in disease risk estimation

For prediction of disease risk there is vast literature on how to build and evaluate risk prediction scores using statistical tools and machine learning, see for instance Ankerst et al. [7, 8], Do et al. [41], Kang et al. [93], Kotze et al. [99], Kruppa et al. [104]. In this area certain challenges arise, where methodological and medical gaps have to be filled. Several of these will be treated in this thesis and are described as follows.

Missing data

As in almost every area where data is measured and to be analyzed, the situation of missing data values can arise. So does this in clinical and epidemiological studies. When risk models should be built by incorporating the multivariate structure of the predictor variables then the problem of missing data is especially difficult; generally all variable values per observation or individual have to be available at once, since variables usually cannot be taken into account sequentially. For situations where this is not the case, solutions have been provided using so-called *multiple imputation* strategies [30] to impute the missing data (see Section 2.2 for types of missing data and imputation techniques). These have been successfully used in risk prediction [98]. However, handling missing values is getting difficult and disputed when the values are missing systematically. In Chapter 5 a special missing data structure will be discussed and a novel strategy will be proposed.

Sample selection bias

So-called *sample selection bias* can be seen as a special case of missing data. It arises when observations in the given sample were not taken at random. In epidemiology certain systematic sampling designs are often used in order to enrich informative observations [49]. The resulting sample selection bias is then corrected for by applying appropriate weighting of observations and adjusting variance based on the design. The methodology there is well-developed for classical statistical analyses; however, there are knowledge gaps in applying modern methods as machine learning algorithms. Chapter 3 will investigate such a situation and develop novel approaches applicable for arbitrary machine learning algorithms. We will examine these on real data by using a biomedical data set for predicting the risk of hepatitis. Chapter 4 is an applied project regarding this topic as the provided data is biased the described way. Correction approaches for training and testing on such data will be applied and extended.

High-dimensional data

A big issue and well-disputed data situation nowadays is high dimensionality. The age of *big data* has arisen several years ago and the expression itself can be interpreted in many ways. However, whenever the *big data* expression is disputed then the interpretation of big data as large data regarding either the number of observations or of the variables is always included in the considerations. The latter issue — having a high number of variables at hand — is surely the less common situation; looking across all kind of disciplines, however, it does not uncommonly occur in medical applications. There, so-called *omics* are common candidates of containing many more variables than observations. These are biological data types typically ending in “-ome” or “-omics” when referring to the actual field of study. Some examples are the genomics studying the lowest functional layer in a biological system, i.e. the genes, transcriptomics studying the RNA, the proteomics studying proteins, or metabolomics studying the chemical processes involving the end products of metabolism. Genome-wide data plays an extraordinary role as the the number of variables is enormous, reaching from hundreds of thousands to millions of SNPs. Handling such kind of data is now commonly done by *genome-wide association studies* (GWASs): in a GWAS DNA sequence variations in terms of SNPs are measured and analyzed on the whole human genome in order to identify risk factors for common diseases [28]. Even though this approach can be incorporated for building risk scores [191], such approaches may not be optimal as, for instance, relatedness of the variables is not incorporated in such a model approach: the term *linkage disequilibrium* (LD) describes “the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP within a population” [28] and is usually high for many positions on the genome. This leads to a scenario of high collinearity in the context of linear regression analysis. Therefore, multivariable statistical learning approaches handling such issues have been applied and compared on such kind of data [165, 189] often involving regularization (see Section 2.3.3), but optimal methods have to be investigated and established for each application in terms of disease type and complexity and species, as those vary in trait architecture, marker density and relatedness of SNPs [165]. Chapter 4 does so for humans in the case of childhood asthma, applying sparse models, i.e. models where the final predictor consists of tens to hundreds of variables, as well as dense models, i.e. models with up to hundreds of thousands of variables in the final predictor. Chapter 5 treats a further case of high-dimensional data, including expression data as high-dimensional predictor variables amongst other things. There, however, the high-dimensionality is less extreme and another issue arises, described as follows.

Multi-omics data

Even though big amounts of predictor variables, such as high-throughput molecular data have been used for developing prediction models for disease outcome for nearly 20 years [67], it still often difficult to combine these with variables of other types of data, which we call modalities [21]. Many diseases are complex and several modalities have to be taken into account for optimal prediction: as previously mentioned, for the example of asthma, these can be genetics, family history, environment, and demographic variables, for instance. Omics data sets as introduced above can represent modalities as well; therefore, the two terms will be treated as equivalent in this thesis. Integration of different modalities have been taken into account in the past: mostly integration of such multiple omics data sets have been focused on analyzing correlation structure [186] but also building risk prediction models has already led to successful prediction [1, 176]. This way of multi-omics integration is sometimes also referred to as *multi-view learning*, especially in the context of machine learning [192]. As typically the dimensionality of the single modalities is very different, those with higher dimension may be preferred by a statistical prediction model than those with lower dimension. Approaches taking this into account have been proposed, for example by analyzing each modality on its own and merging them or, in amended form, merging them at different stages of the analysis [62, 195]. Another approach, the IPF-LASSO, incorporates all modalities simultaneously with an additional optimized weighting of each modality [20]. We will use this approach in Chapter 4. However, when the number of modalities gets too high, this approach is too unstable. Another issue arises when not all modalities are measured for all observations and only few complete observations are available. The latter two issues come up in the data analyzed in Chapter 5: there, immunological modalities in addition to genetic and environmental ones have to be taken into account. We will propose a model strategy in order to tackle both issues.

1.4 Objectives of the thesis

The overall aim of this thesis — motivated by the application on studies on childhood asthma — was to improve disease risk prediction for complex data structures, such as general and complex missing data, data suffering from sample selection bias, large and high-dimensional data, and multi-omics data.

In particular, we aimed at developing new methodologies and strategies and sought for selection and application of most appropriate and modern methods in the field of machine learning in order to exploit techniques for disease risk prediction at the best.

An entailing subsidiary aim was to identify best prediction models as well as the most important factors and variables for the disease childhood asthma.

1.5 Contributing manuscripts

Several parts of this thesis have been published by or submitted to peer reviewed journals. The articles were written in collaboration with co-authors, mostly from the Institute of Computational Biology, Helmholtz Zentrum München, the Technical University of Munich, or from the Dr. von Hauner Children's Hospital. The articles are listed in the following, along with individual contributions of the (main) investigators where relevant for this thesis. The original papers describe the contributions of all authors.

- **Chapter 3:**

[100]: Norbert Krautenbacher, Fabian J. Theis, and Christiane Fuchs. Correcting Classifiers for Sample Selection Bias in Two-Phase Case-Control Studies. *Computational and Mathematical Methods in Medicine*, 2017:18, 2017. doi: 10.1155/2017/7847531. <https://www.hindawi.com/journals/cmmm/2017/7847531/>

I extended or modified the described methods (correction approaches for sample selection bias) which have been proposed in literature and I developed the methods *Stochastic Inverse-Probability Oversampling* and *Parametric Inverse-Probability Bagging*. I designed and conducted the simulation study and implemented all correction approaches. I implemented the R-package *sambia* with help of student assistants Kevin Strauß and Maximilian Mandl. All aspects of the project were supervised by Christiane Fuchs. The manuscript was written in cooperation with Christiane Fuchs. Apart from minor modifications, Chapter 3 and Krautenbacher et al. [100] match. Some parts of Krautenbacher et al. [100] were already covered by the methodological background (Chapter 2) and thus were left out for Chapter 3.

- **Chapter 4:**

[102]: Norbert Krautenbacher, Michael Kabesch, Elisabeth Horak, Charlotte Braun-Fahrländer, Jon Genuit, Andrzej Boznanski, Erika von Mutius, Fabian J Theis, Christiane Fuchs, and Markus J Ege. Predicting childhood asthma risk by genetic and environmental variables. *submitted*, 2018

I initiated and conducted the methodology used in the paper. I did all statistical analyses and built and implemented the *confidence intervals for AUC by bootstrap*

using selection probabilities corrected for sample selection bias and the bootstrap test for pairwise AUC comparison using selection probabilities. I modified several built-in R methods to be applicable for appropriate weighting of observations (from packages *ipflasso*, *roc*). All aspects regarding methodology and analyses were supervised by Christiane Fuchs. Medical details for the manuscript were provided by Markus Ege. The manuscript was written in cooperation with Markus Ege and Christiane Fuchs. Chapter 4 and Krautenbacher et al. [102] match apart from minor changes; however, parts from the supplement of the latter were included in the main text of this thesis's chapter or left out for Chapter 4 and already included in the methodological background (Chapter 2).

- **Chapter 5:**

[101]: Norbert Krautenbacher, Nicolai Flach, Andreas Böck, Kristina Laubhahn, Michael Laimighofer, Fabian J Theis, Donna P Ankerst, Christiane Fuchs, and Bianca Schaub. Classifying childhood asthma phenotypes from genetic, immunological and environmental factors: A strategy for high-dimensional multivariable analysis. *submitted*, 2018

I initiated the methodology used in the paper and conducted the analyses together with the student Nicolai Flach who did most of the implementation tasks. I developed the *Combination strategy on complex multi-omics data structure*. Christiane Fuchs supervised the methodology and analyses. Medical details for the manuscript were provided by Bianca Schaub, Andreas Böck, and Kristina Laubhahn. The manuscript was written in cooperation with Bianca Schaub, Andreas Böck, Kristina Laubhahn, Christiane Fuchs, Donna Ankerst and Michael Laimighofer.

Chapter 5 and Krautenbacher et al. [101] match apart from minor changes. However, parts from the supplement of the latter were included in the main text of this thesis's chapter or left out for Chapter 5 when already included in the methodological background (Chapter 2).

Further contributing manuscripts

I was involved in further research projects on prediction of disease risk on complex data sets which, however, are not contained in the main focus of the thesis.

- [98]: Ivan Kondofersky, Michael Laimighofer, Christoph Kurz, Norbert Krautenbacher, Julia Söllner, Philip Dargatz, Donna Ankerst, and Christiane Fuchs. Three

general concepts to improve risk prediction: good data, wisdom of the crowd, recalibration. *F1000Research*, 5:2671, 2016. doi: 10.12688/f1000research.8680.1

- [70]: Justin Guinney, . . . , Norbert Krautenbacher, . . . , and Yuxin Zhu. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. *The Lancet Oncology*, 18(1): 132–142, feb 2018. ISSN 1470-2045. doi: 10.1016/S1470-2045(16)30560-5. URL [http://dx.doi.org/10.1016/S1470-2045\(16\)30560-5](http://dx.doi.org/10.1016/S1470-2045(16)30560-5) (complete author list: s. Bibliography)
- [159]: Fatemeh Seyednasrollah, . . . , and Prostate Cancer DREAM Challenge Community(. . . , Norbert Krautenbacher, . . .). A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer. *JCO Clinical Cancer Informatics*, (1):1–15, 2017. doi: 10.1200/CCI.17.00018. URL <https://doi.org/10.1200/CCI.17.00018> (complete author list: s. Bibliography)

Chapter 2

Methodological background

This chapter summarizes the methodology that has been used or is required for further chapters. Parts of texts on notation or description of methodology match with those of Krautenbacher et al. [100], Krautenbacher et al. [101], or Krautenbacher et al. [102] with minor modifications.

2.1 General notation

For a data set we let n be the sample size and p be the number of variables (usually of the covariates). We assume a set of observations $\{(\mathbf{x}_i, y_i)\}_{i=1, \dots, n}$ which are drawn independently from a distribution D . The domain of D is $\mathcal{X} \times \mathcal{Y}$ with \mathcal{X} the feature space and \mathcal{Y} a measurable space. Throughout the thesis, \mathcal{Y} is a discrete label space since we focus on classifiers in this work. In the case of a binary label space Y is coded by $\{0, 1\}$ — in this work this usually refers to the disease status: “disease ($Y = 1$) vs. no disease ($Y = 0$)”. The setting can be extended to more than two outcome categories, for instance when there are several subtypes of the disease. We will denote random variables by capital letters and realizations (i. e. observations in the sample) by lower-case letters. Thus, the matrix \mathbf{x} contains all covariates of the given data, we let \mathbf{x}_i be the values for subject i and x_{ij} be the value for subject i and covariate (feature) j , unless otherwise stated. \mathbf{X} is then the corresponding random variable.

2.2 Imputation of missing data

Missing data occur in many practical data applications, in questionnaire data, for instance, and thus requires appropriate solutions avoiding bias as far as possible.

In this chapter we will denote $\mathbf{X}^{(*)}$ as a multi-dimensional random variable with missing values, $\mathbf{X}_{obs}^{(*)}$ as the observed values of $\mathbf{X}^{(*)}$ (in the sense of non-missing values rather than of realizations) and $\mathbf{X}_{mis}^{(*)}$ as the missing values of $\mathbf{X}^{(*)}$. Thus, realizations $\mathbf{x}_{obs}^{(*)}$ and $\mathbf{x}_{mis}^{(*)}$ taken together contain all values of the data $\mathbf{x}^{(*)}$, the part $\mathbf{x}_{mis}^{(*)}$, however, is unknown. $S \in \mathcal{S}$ with \mathcal{S} a binary space denotes the variable that indicates whether $\mathbf{X}^{(*)}$ is observed: for a variable j , $S_j = 1$ if $X_j^{(*)}$ is observed (i.e. $X_j^{(*)}$ is a subset of $\mathbf{X}_{obs}^{(*)}$), $S_j = 0$ if $X_j^{(*)}$ is not observed (i.e. $X_j^{(*)}$ is a subset of $\mathbf{X}_{mis}^{(*)}$). The notation and descriptions in this section are partly similar to Buuren [30].

2.2.1 Types of missing data

According to Rubin [149] data can be missing in three different ways: *data missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). The following description is close that of Buuren [30].

MCAR is the missing data case causing the least problems when data has to be imputed, it occurs when the probability of values being missing is the same for all observations. Thus, in such a case, the cause of the missingness in the data is unrelated to the data itself. Denoting ψ as the parameters of the missing data model formally data is MCAR if

$$P(S = 0 | \mathbf{X}_{obs}^{(*)}, \mathbf{X}_{mis}^{(*)}, \psi) = P(S = 0 | \psi).$$

For instance, taking a random sample from a population where each observation has equal probability of getting into the sample leads to missingness completely at random: observations not included in the resulting sample are then MCAR.

Data missing at random (MAR) is similar to MCAR; however, the probability of missingness depends on *observed* data. Again, as an example, a sample can be taken from a population where the probability to be included depends on a known property. Then data is MAR if

$$P(S = 0 | \mathbf{X}_{obs}^{(*)}, \mathbf{X}_{mis}^{(*)}, \psi) = P(S = 0 | \mathbf{X}_{obs}^{(*)}, \psi).$$

The most complex missingness situation is when data are missing not at random (MNAR). This occurs when neither MCAR nor MAR can be assumed, that is the probability of the missingness depends on unobserved data. Again, the example of taking a sample from

a population can be applied: the probability to be included depends on an unknown property. In a survey, income can be one variable with missings, where covariates which are related to income are not given, for instance. The missing data model

$$P(S = 0 | \mathbf{X}_{obs}^{(*)}, \mathbf{X}_{mis}^{(*)}, \psi)$$

cannot be simplified in the case of MNAR.

In general, imputation for this type of missing data is difficult. The methods described in this section are applicable for situations of MCAR or MAR. For handling MNAR, strategies for finding more data on the explanations of the missingness should be found. The problem of MNAR is further discussed in McPherson et al. [118], for instance.

2.2.2 Single imputation techniques

Many techniques for handling missing data have been developed. The probably simplest one is the usage of only complete observations, i.e. observations containing missing values at any variable are simply deleted. This approach is easy to apply and unbiased under MCAR [30]. However, it is wasteful because a lot of given information is not used and estimates can get imprecise.

Therefore, many commonly used techniques for handling missing data are based on imputation methods replacing each missing value by one estimated value. This is called *single imputation*. However, also many of these methods have disadvantages: the simplest technique amongst those may be mean imputation, i.e. a missing value per variable is simply replaced by the variable's mean, so $\hat{x}_{mis,j}^{(*)} = \frac{1}{n} \sum_i x_{obs,ij}^{(*)}$; it is unbiased for the mean under MCAR, but disturbs the distribution of the data, and underestimates their covariances [30]. More advanced methods are regression imputation or stochastic imputation: only complete observations are used for building a regression model with choosing the variable as response which missing values should be imputed for, i.e. $\mathbf{X}_{obs}^{(*)}$. The model fit is then used for predicting this variable's missing values $\mathbf{X}_{mis}^{(*)}$ by replacing the missing values by the predicted ones (regression imputation). One can add some appropriate noise to the prediction in order to reflect the uncertainty in the data which is underestimated by simple regression imputation (stochastic imputation). Both approaches, however, still do not take uncertainty into account appropriately, as imputed data are treated the same way as the given data values. Generally all these approaches lead to underestimation of the standard errors in the subsequent statistical analyses [30]. Figure 2.1 illustrates single imputation techniques and their disadvantages.

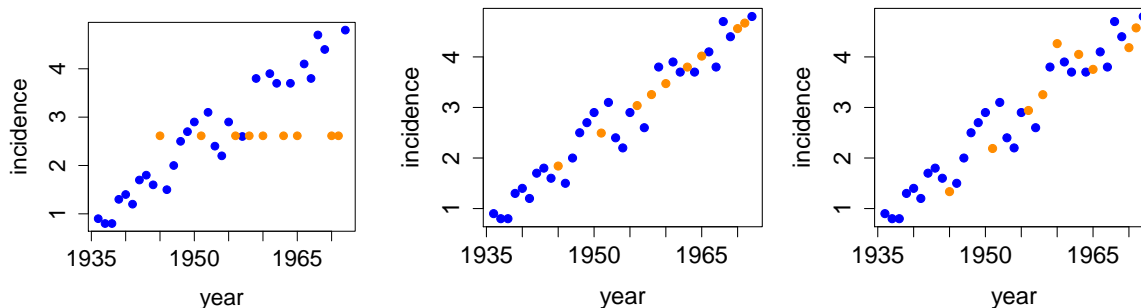


Figure 2.1: Single imputation on variables *year* and *incidence* for melanoma skin cancer (data set *melanoma* from R-package *lattice* [152]). Blue indicates observed data points, orange indicates imputed values. left figure: mean imputation. The distribution gets disturbed, the variance is underestimated and the correlation is biased to zero. center figure: regression imputation. Correlation between the two variables is increased artificially and the variance is underestimated systematically. Right figure: stochastic imputation. Uncertainty is not taken into account as imputed data is treated as observed data.

2.2.3 Multiple imputation

A solution to the issues described in the previous section is *multiple imputation*. It generally contains the following steps introduced by Rubin [149]: the given incomplete data set is imputed by a technique leading to several imputed versions of the data set. The desired statistical analyses are conducted on each of the imputed data sets separately. In a final step the results are aggregated (pooled) to one result with appropriate unbiased standard errors (cf. Buuren [30] for more details).

For multiple imputation, similar techniques as in single imputation can be used. However, at least one difference is required, when $\mathbf{X}^{(*)}$ should be imputed: the integrated single imputation techniques must lead to varying imputation values each time they are applied.

One variable with missing values

As described in the previous section, missing values $\mathbf{X}_{mis}^{(*)}$ can be predicted from using the remaining variables in a linear regression setting (regression imputation). Since this method does not take into account the uncertainty of the model, i.e. only the value which is most likely is used, added noise to the predicted values solves this problem partly (stochastic imputation). Therefore values, for instance from a normal distribution with

zero mean and variance of the residuals of the regression model, are taken and added to the predicted value. As the true parameters of the regression model are usually not known, parameter uncertainty can be taken into account in addition. To do so, there are two main methods: one can draw the parameters from their posterior distributions in a Bayesian fashion or alternatively apply bootstrapping, i.e. observed data is resampled with replacement and parameters are re-estimated from the resulting samples.

A further alternative to these two techniques uses similar ideas, but exclusively imputes missing values by observed values, namely *predictive mean matching* (PMM) [112]. The method uses the Bayesian regression technique; however, here when potential values $\mathbf{X}_{mis}^{(*)}$ are predicted, a set of values of $\mathbf{X}_{obs}^{(*)}$ which are close to the predicted values are identified. From this set one observation is taken randomly in order to substitute the missing value. This method has several advantages apart from being robust and easy to use. For instance, it guarantees that imputed observations are realistic and imputed values outside of the observed range of the variable to be imputed are impossible to occur. In all situations of this thesis where multiple imputation is involved we choose PMM for imputing continuous values.

Until now, we only described techniques for the situation where $\mathbf{X}^{(*)}$ takes continuous values. Alternatives for categorical variables exist and partly work analogously, but involve logistic regression (see Section 2.3.1) or multinomial regression (see Section 2.3.2) instead of linear regression, for instance [30].

More than one variable with missings

Until now, we have described solutions for imputing in cases where one variable in a data set has missing values. In practice, usually more than one variable contain missing values which makes it more difficult to apply these imputation techniques. There are generally several principles of handling the situation. For instance, data can be imputed *monotonically*, i.e. values are imputed by a sequence of univariate methods. This strategy, however, is restricted to a certain missingness pattern. A strategy applicable to more general missingness patterns specifies a multivariate model by a set of conditional univariate models. By iterating conditional models values are drawn in order to get imputed values. This is known as *fully conditional specification* or *chained equations*.

The latter strategy has shown several advantages, it is flexible and easy to apply and has shown satisfactory results (unbiased estimates) in simulations and practice [23, 81, 111, 137, 140]. Therefore, we employ the chained equations strategy, using the *multiple imputation by chained equations* (MICE) algorithm and the corresponding R package [173] in

the two more applied projects of the thesis where missing data occurs (Chapters 4 and 5). The MICE algorithm is initiated by drawing values randomly from observed data and the incomplete data is then imputed variable by variable. Then several iterations are performed using a Markov chain Monte Carlo method (MCMC).

In detail, for building an imputation data set using T iterations for data consisting of p variables, the MICE algorithm works as follows (similar in Buuren [30]).

1. Specify $P(X_{mis,j}^{(*)}|X_{obs,j}^{(*)}, \mathbf{X}_{-j}^{(*)}, S)$ as an imputation model for variable $X_j^{(*)}$ where $\mathbf{X}_{-j}^{(*)}$ is $\mathbf{X}^{(*)}$ without $X_j^{(*)}$
2. For each j , impute $X_{mis,j}^{(*)}$ by starting imputations with random draws from $X_{obs,j}^{(*)}$
3. For t in $1, 2, \dots, T$ do
 - For j in $1, 2, \dots, p$ do
 - (a) Let $\mathbf{X}_{-j}^t = (X_1^t, \dots, X_{j-1}^t, X_{j+1}^{t-1}, \dots, X_p^{t-1})$ be the currently complete data without $X_j^{(*)}$
 - (b) Draw $\dot{\phi}_j^t \sim P(\phi_j^t | X_{obs,j}^{(*)}, \mathbf{X}_{-j}^t, S)$
 - (c) Draw imputations $X_j^t \sim P(X_{mis,j}^{(*)} | X_{obs,j}^{(*)}, \mathbf{X}_{-j}^t, S, \dot{\phi}_j^t)$
4. Obtain complete data $\mathbf{X} := \mathbf{X}^T$

ϕ_j^t denotes the unknown parameter of the imputation model for variable p and iteration t ; $\dot{\phi}_j^t$ represents a value randomly drawn from the posterior distribution of ϕ_j^t .

We portray this algorithm with the following example (similar in Azur et al. [12]): We assume a data set of 3 variables, which are age, income, and gender; each of them contains missing values. The MICE algorithm first specifies the imputation model (step 1 of the algorithm) and each variable is imputed with a starting imputation by filling the missings with random draws from the values observed (step 2). Starting with age, the originally missing values are set back to missings (step 3(a)) and only income and gender are used for building an imputation model (step 3(b)) in order to draw imputations for age (step 3(c)). Steps 3(a) to (c) are carried out for income, i.e. originally missing values are set back to missings and predicted by an imputation model using age and gender. The procedure is analogous for gender. Iterating these steps (3(a) - (c)) is then repeated several times (step 3, T times) until convergence. The final set of imputed values together with the observed values build one complete data set (step 4).

Markov chains have to fulfill certain properties so they converge to a stationary distribution

[158]. First, a Markov chain has to be irreducible, i.e. the chain should be able to reach all possible states from any state. This property is fulfilled for the MICE algorithm. Second, a Markov chain has to be aperiodic: there should be no oscillation between the different chains. Aperiodicity is not necessarily guaranteed for MICE, but can be diagnosed and avoided by stopping the chain at different points or by addition of noise. Third, a Markov Chain has to be recurrent: a state is recurrent if it is not transient which is the case if, given we start at a certain state, the probability of never returning to this state is greater than zero. This property is not directly satisfied for the MICE algorithm but again can be diagnosed. According to Buuren [30], in practical applications on data non-recurrence is usually mild or even absent.

2.3 Statistical and machine learning models

With the following subsections we will give a brief overview of some classical statistical models and modern machine learning algorithms. Each of the described methods will at least once be used in the further chapters.

Statistical learning models (classifiers) are algorithms which learn on given training data in order to predict an outcome variable as well as possible. A classification algorithm as a special case predicts categorical outcome and thus is eligible for predicting (disease) phenotypes or the risk of those. A classifier can be defined as

$$\varphi : \begin{cases} (\mathcal{X} \times \mathcal{Y})^{\times n} \times \mathcal{X} & \rightarrow \mathcal{Y}^* \\ ((\mathbf{x}, \mathbf{y}), \mathbf{X}) & \mapsto \varphi((\mathbf{x}, \mathbf{y}); \mathbf{X}), \end{cases}$$

where the given learning data set $(\mathbf{x}, \mathbf{y}) = ((\mathbf{x}'_1, y_1)', \dots, (\mathbf{x}'_n, y_n)')$ is mapped to the prediction (in our case classification) rule and applied to the random variable \mathbf{X} . In classical regression, φ predicts a value for Y , and hence $\mathcal{Y}^* = \mathcal{Y}$. In classification, \mathcal{Y}^* is often a “continuous pendant” of \mathcal{Y} . In the logistic regression model (see next subsection), for instance, one models $P(Y = 1) \in \mathcal{Y}^* = [0, 1]$ instead of $Y \in \mathcal{Y} = \{0, 1\}$.

In Chapters 4 and 5 we use notation in more applied manner. Since we have given test data at hand we write in a less general context and apply the classifier to the new data \mathbf{x}_{new} instead of the random variable \mathbf{X} .

In the following sections we will at first introduce models based on generalized linear regression (Sections 2.3.1 to 2.3.4). There the general goal is to minimize the expectation

of a loss $L(Y, \varphi)$, i.e.

$$\hat{Y} = \arg \min_{\varphi((\mathbf{x}, y); \mathbf{X})} \mathbb{E}_{\mathbf{x}, y} L(Y, \varphi((\mathbf{x}, y); \mathbf{X})). \quad (2.1)$$

2.3.1 Logistic regression

We employ logistic regression [51] as a common classical binary classification method. It is a common way to build a prediction model, for instance for a disease, and thus a risk score. The model assumes $Y|\mathbf{X}$ to be Bernoulli distributed with success probability

$$P(Y = 1|\mathbf{X}) = (1 + \exp\{-(\beta_0 + \mathbf{X}'\boldsymbol{\beta})\})^{-1},$$

where β_0 and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ are unknown parameters representing the effects of the features \mathbf{X} on the outcome variable Y . Here, the output of prediction model $\varphi((\mathbf{x}, y); \mathbf{X})$ corresponds to the probability $P(Y = 1|\mathbf{X})$.

In order to obtain the estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ from the data, one performs maximum likelihood estimation by maximizing the log-likelihood

$$\ell(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^n (y_i(\beta_0 + \mathbf{x}'_i\boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \mathbf{x}'_i\boldsymbol{\beta})))$$

with respect to β_0 and $\boldsymbol{\beta}$.

For inferring the parameters solving the maximum likelihood estimation equation is not possible in closed form here. Thus, iterative numerical algorithms can be employed in order to calculate $(\beta_0, \boldsymbol{\beta})$ as the zeros of the derivative of $\ell(\beta_0, \boldsymbol{\beta})$. This can be done by the Fisher-Scoring algorithm which here is equivalent to Newton's method (for details see Fahrmeir et al. [51], for instance).

2.3.2 Multinomial regression

The multinomial regression generalizes the logistic regression model: it extends the model from $K = 2$ to $K > 2$ outcome categories. This can be modeled by

$$P(Y = l|\mathbf{X}) = \frac{\exp\{\beta_{0l} + \sum_{j=1}^p X_j\beta_{jl}\}}{\sum_{k \in \{1, \dots, K\}} \exp\{\beta_{0k} + \sum_{j=1}^p X_j\beta_{jk}\}}$$

with $l = 1, \dots, K-1$ which corresponds to a more symmetric approach regarding parametrization than a traditional approach [57]. X_j refers to the j -th covariate. Thus, for covariate j there are K parameter estimates β_{jk} instead of one β_j . The parameter estimation steps will be done for each l .

2.3.3 Penalized regression

The use of multivariable logistic or multinomial regression is suitable for classification problems; however, it can be improved in prediction by using regularization [76]. Further, if $n > p$, regression without regularization is analytically not feasible. In regularized regression parameters are estimated by

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg \min_{(\beta_0, \boldsymbol{\beta})} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \cdot \text{pen}(\boldsymbol{\beta}) \right\}$$

with λ as tuning parameter and pen a penalization function. In logistic or multinomial regression the term $\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2$ is replaced by $-\ell(\beta_0, \boldsymbol{\beta})$ with ℓ denoting the log-likelihood function. The penalization function can be chosen to be based on the L_1 -norm, i.e.

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^p |\beta_j|$$

which corresponds to the *LASSO* penalty. It reduces the dimension of the data and performs hard thresholding by setting coefficients of non-predictive or strongly correlated variables to zero. Thus the LASSO performs variable selection. This is desired in many data applications where it is unknown whether or which variables in a regression setting are predictive and which variables are just noise or lead to collinearity. Therefore LASSO is most appropriate when there is a small to moderate number of effects with moderate size [168].

Another option for the penalty function is the L_2 -norm, i.e.

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$$

which corresponds to the *ridge* penalty. It shrinks the coefficients of correlated variables towards zero but without reaching the zero [76]. This version may be preferable when there are many small effects rather than few moderate or large effects in the covariates [168].

Both penalties can be included in the penalization function, i.e.

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|)$$

which yields the *elastic-net* penalty. It forms a compromise between the LASSO penalty and the ridge penalty. Elastic net performs variable selection as the LASSO does but often leads to better performance. It is especially useful when p is much larger than n where LASSO tends to perform less successful [87].

2.3.4 Integrative L_1 -penalized regression with penalty factors based on multi-omics data

As described in the introduction, a further issue in medical data can be the situation of having several modalities. As previously mentioned, by modalities we refer to blocks of variables which biologically belong together, for instance variables containing information about the environment, demographics, family history or genetics. If these differ substantially in their dimensions, practical problems can occur, i.e. variables of low-dimensional modalities can get put at a disadvantage [21]. Therefore there is a modified version of the LASSO, namely the IPF-LASSO (integrative L_1 -penalized regression with penalty factors), assigning additionally different penalty factors to different data modalities [21]. Their coefficients are obtained by minimizing

$$\sum_{i=1}^n \left(y_i - \sum_{m=1}^M \sum_{j=1}^{p_m} x_{ij}^{(m)} \beta_j^{(m)} \right)^2 + \text{pen}^{(m)}(\boldsymbol{\beta})$$

where

$$\text{pen}^{(m)}(\boldsymbol{\beta}) = \sum_{m=1}^M \lambda_m |\boldsymbol{\beta}^{(m)}|$$

with p_m the number of variables from modality m ($m = 1, \dots, M$), $\lambda_m > 0$ the tuning parameter applied to the variables from modality m ($m = 1, \dots, M$) which are denoted as $X_1^{(m)}, \dots, X_{p_m}^{(m)}$ with their realizations $x_{i1}^{(m)}, \dots, x_{ip_m}^{(m)}$, and $\boldsymbol{\beta}^{(m)} = (\beta_1^{(m)}, \dots, \beta_{p_m}^{(m)})$. As it will be required in Chapter 4, this framework can be adjusted to logistic regression analogously to the previous section. Here, the complexity for parameter estimation is not higher than for the LASSO, i.e. standard LASSO algorithms can be used [21].

The IPF-LASSO has similarities to group or sparse group LASSO, but is more flexible

regarding the variation of the $L1$ shrinkage parameters for the different modalities. Therefore, however, more tuning parameters have to be estimated.

2.3.5 Classification trees

Here we will introduce a model conceptually different to the previously described methods. Classification trees (or regression trees for continuous outcome variables) [76] segment the predictor space into regions. In these regions constant functions are assumed and the segmentation can be interpreted as a tree. For an example see Figure 2.2.

For a p -dimensional feature space let V be the number of regions R_1, \dots, R_V . Then the prediction model of a tree is given as

$$\varphi((\mathbf{x}, y); \mathbf{X}) = \sum_{v=1}^V c_v I(\mathbf{X} \in R_v)$$

where I denotes the indicator function. In classification c_v is defined as

$$c_v := \operatorname{argmax}_v p_{vl}$$

where for outcome categories $l = 1, \dots, K$

$$\hat{p}_{vl} = \frac{1}{|\{\mathbf{x}_i | \mathbf{x}_i \in R_v\}|} \sum_{\mathbf{x}_i \in R_v} I_{y_i=l}$$

which is the percentage of observations of category l in region R_v . Thus, the most probable class in the region is selected. Trees are grown in a greedy algorithmic fashion by iteratively splitting the feature space along a certain splitting variable with a certain splitting value by minimizing a cost function. This procedure is repeated recursively until some depth is reached and all regions are built. For details, see for example Hastie et al. [76].

The following two sections describe so-called *ensemble learners*. The principle of ensemble learning is to improve predictive performance by combining a collection of base learning models to one powerful prediction model [76].

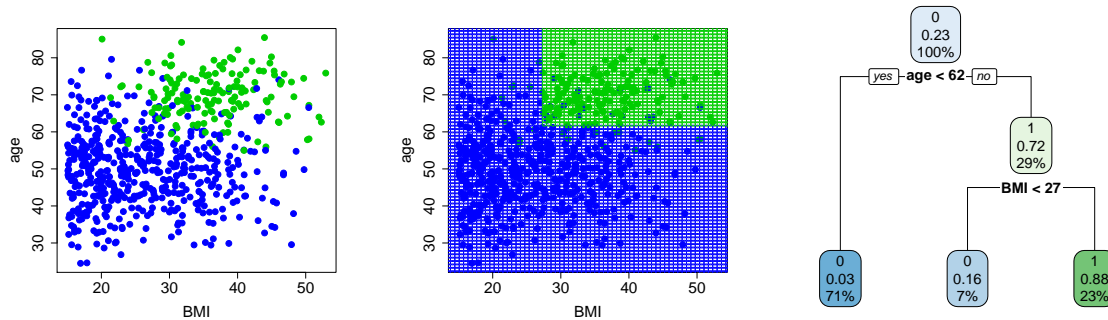


Figure 2.2: Segmentation of a two-dimensional feature space (toy data with variables *age* and *BMI*) into regions by a classification tree with two splits. Left figure: data with two outcome categories, diseased (green) and healthy controls (blue). Center figure: regions classifying observations as diseased (green area) or healthy (blue area). Right figure: Resulting tree corresponding to center figure.

2.3.6 Random forest

Random forest is a famous machine learning method. It is based on regression or classification trees.

Prediction model

Random forests are ensembles of decision trees and a modification of bagging [26]. The basic procedure of the learning algorithm is the following:

1. A bootstrap sample is drawn from the given learning data set.
2. A decision tree is grown by constructing recursive binary splits to the given data based on the features (see Section 2.3.5).
3. At each node only a subset of features is selected at random.
4. Steps 1 to 3 are repeated and all trees are averaged; class probabilities can be estimated as the relative frequency of the class of interest for a terminal node.

An essential step which is different from common bagging (cf. Section 3.2.1) is Step 3. The random selection of features de-correlates the trees and makes the bagging procedure more efficient.

Variable importance

Contrary to generalized linear regression methods where the influence of a variable can easily be interpreted by its effect β_i , random forests are less easy to interpret. However, the most important variables can be determined, for instance by a standard measure of variable importance, the permutation importance. For $j \in \{1, \dots, p\}$, it is given by

$$VI_j = \frac{1}{\text{ntree}} \sum_{t \in \{1, \dots, \text{ntree}\}} \frac{1}{|\text{OOB}_t|} \sum_{i \in \text{OOB}_t} \{I(y_i \neq \hat{y}_{it}^*) - I(y_i \neq \hat{y}_{it})\},$$

where ntree denotes the number of trees in the forest, I denotes the indicator function, OOB_t the set of indices for observations not selected for building tree t (out-of-the-bag observations), \hat{y}_{it} the predictions by the t -th tree before, and \hat{y}_{it}^* after permuting the j -th variable's values. Further details can be found in Janitza et al. [89], for instance. In order to obtain a selection of predictive variables, we applied a non-parametric version of the permutation-based test of Altmann et al. [4]: after VI_j for the j -th variable has been calculated, a distribution of null importance values is built repeating three steps s times:

- (i) Permuting the values of the response y
- (ii) Fitting a new random forest using the permuted response
- (iii) Computing VI_j again

The p-value is computed as the fraction of null importance values which exceed the originally computed importance.

2.3.7 Gradient boosting

The following description of gradient boosting adjusted to our notation is similar to e.g. Hastie et al. [76] (p.359 et seq.) or Ridgeway [142].

Boosting is — as the random forest — an ensemble learner; it combines many weak learners to a strong learner in a — contrary to random forests — stage-wise way by optimizing a certain loss function. Not only because of different choices of this loss functions there are various different versions of boosting. Here we focus on a version which is a powerful prediction tool: Friedman's gradient boosting machine [58] and its extension to stochastic

gradient boosting [59]. Gradient boosting outperforms other learning algorithms in many applications, for instance in Ogutu et al. [132]. It is applicable when n is much larger than p , addresses multicollinearity problems, and optimizes the prediction accuracy; thus it is a efficient prediction model suitable for predicting disease risk.

Prediction model

Similar to the regression setting, in gradient boosting we try to find an estimator that minimizes the expectation of our loss function $L(Y, \varphi)$, see Equation 2.1. The loss function is basically arbitrary and can be chosen to be the squared error loss in linear regression or the misclassification rate in classification problems, for instance. In practice, when a set of realizations $\{(\mathbf{x}_i, y_i)\}_{i=1, \dots, n}$ is given one tries to minimize the empirical risk

$$\mathcal{R} = \frac{1}{n} \sum_{i=1}^n L(y_i, \varphi((\mathbf{x}, y); \mathbf{x}_i))$$

with respect to φ and its parameters. The idea of gradient boosting is to do so by modifying the current estimate \hat{Y} by adding a further function in a greedy algorithmic way: the negative gradient of

$$\mathcal{J}(\varphi) = \sum_{i=1}^n L(y_i, \varphi((\mathbf{x}, y); \mathbf{x}_i))$$

is calculated in each iteration — in case of the mean squared loss these are the residuals — and the estimate \hat{Y} can be updated per iteration by

$$\hat{Y} \leftarrow \hat{Y} - \rho \nabla \mathcal{J}(\varphi)$$

where ρ corresponds to the step size along the direction of the greatest descent in a gradient descent algorithm. With this procedure, however, prediction on an independent validation set would not be successful and one further step is important: Friedman proposes to estimate the negative gradient by using covariate information in a base-learning procedure in order to approximate the gradient, usually — and as in our applications — a regression tree. The final gradient boosting algorithm is in detail described by the following steps:

1. Initialize $\varphi_0((\mathbf{x}, y); \mathbf{X}) = \arg \min_{\rho} \sum_{i=1}^n L(y_i, \rho)$
2. For t in $1, 2, \dots, T$ do

(a) For i in $1, 2, \dots, n$ compute

$$z_i = \left[\frac{\partial}{\partial \varphi((\mathbf{x}, y); \mathbf{x}_i)} L(y_i, \varphi((\mathbf{x}, y); \mathbf{x}_i)) \right]_{\varphi=\varphi_{t-1}}$$

(b) Fit a learner $\Gamma_t((\mathbf{x}, z); \mathbf{x}_i)$ predicting z_i for all i .

(c) Compute the gradient step size as

$$\rho = \arg \min_{\rho} \sum_{i=1}^n L(y_i, \varphi_{t-1}((\mathbf{x}, y); \mathbf{x}_i) + \rho \Gamma_t((\mathbf{x}, y); \mathbf{x}_i))$$

(d) Update

$$\varphi_t((\mathbf{x}, y); \mathbf{X}) = \varphi_{t-1}((\mathbf{x}, y); \mathbf{X}) + \rho \Gamma_t((\mathbf{x}, y); \mathbf{X})$$

3. Output $\hat{Y} = \varphi_T((\mathbf{x}, y); \mathbf{X})$

Several extensions improve the framework of this algorithm as proposed in Friedman [58] and Friedman [59]. In particular, Friedman [59] proposes a further improvement which he calls *stochastic* gradient boosting algorithm: observations are sampled uniformly without replacement before each gradient step. This leads to variance reduction and led to improved performance of the overall prediction model.

Variable importance

Friedman [58] suggests a variable importance measure similar to the one for random forest. For boosted tree based methods an approximate *relative influence* measure of variable X_j by Breiman et al. [24] is used which is defined as

$$\widehat{\text{RI}}_j(\text{tree}) = \sum_{s=1}^{S-1} \hat{i}_s^2 I(v_s = j)$$

for a tree with S terminal nodes, index s running through all non-terminal nodes (since a tree with S terminal nodes has $S - 1$ non-terminal nodes), v_s the splitting variable which is associated with s , I the indicator function, and \hat{i}_s^2 the empirical improvement when v_s is split. Friedman [58] then calculates the relative influence over all T trees which have been generated by the boosting algorithm, that is

$$\widehat{\text{RI}} = \frac{1}{T} \sum_{t=1}^T \widehat{\text{RI}}_j(\text{tree}_t).$$

2.3.8 Naive Bayes

The naive Bayes classifier is another common machine learning algorithm for classification (see e.g. Hastie et al. [76], p. 210 et seq.). It is simple and a computationally fast algorithm and works well in high dimensions, especially if the naive assumption is nearly met; it assumes independence between the p features and simply calculates for each class j that can be attained by Y the marginal classifier

$$\varphi^{(j)}((\mathbf{x}, y); \mathbf{X}) = \prod_{k=1}^p \varphi^{(j,k)}((\mathbf{x}, y); X^{(k)})$$

by estimating feature-wise classifiers $\varphi^{(j,k)}$ via one-dimensional kernel-density estimation. That means, the impact of each feature $X^{(k)}$ is estimated separately and combined to an overall classifier.

2.4 Validating prediction models

In general, there is no prediction model that outperforms all other prediction models in every problem; this is the general statement of the “no free lunch theorem” [187]. Therefore on some given data different models are typically applied and validated in order to assess and compare how well they perform, so that the best model can be identified.

2.4.1 Measuring performance

The prediction accuracy can be evaluated by the ability of how well a prediction model can classify correctly, e.g. how well it can differentiate whether a subject would suffer from a disease or not. One calculates the receiver-operator-characteristics (ROC) curve, which is a measure widely used for evaluating prediction models when the disease response variable is binary [54]. Since our prediction model returns values between 0 and 1 corresponding to the probability of a subject having the disease, a decision whether the subject would be diseased can be made by choosing a threshold, say c . If the threshold is exceeded, the subject is labeled as being tested positive, otherwise as being tested negative. If pr

denotes the predicted risk, there are two commonly used measures of correct prediction: the true positive rate or sensitivity is defined by

$$\text{TPR}(c) = P(\text{pr} \geq c | \text{diseased})$$

and the false positive rate which is $1 - \text{specificity}$ is defined by

$$\text{FPR}(c) = P(\text{pr} \geq c | \text{not diseased}).$$

Displaying sensitivity (TPR) against $1 - \text{specificity}$ (FPR) for all possible choices of c yields the ROC curve

$$\text{ROC}(\cdot) = \{\text{FPR}(c), \text{TPR}(c), c \in (-\infty, +\infty)\} = \{(t, \text{ROC}(t)), t \in (0, 1)\}$$

where in the latter expression ROC is the function mapping t to $\text{TPR}(c)$ and $\text{FPR}(c) = t$. The area under the ROC curve (AUC) is a common measure for model comparison and evaluation [31, 75] and defined as

$$\text{AUC} = \int_0^1 \text{ROC}(t) dt.$$

The AUC can also be retained by calculating the Wilcoxon test of ranks [75]; it can be seen as the probability that a classifier will rank a randomly chosen observation with value 1 higher than a randomly chosen observation with value 0 [54]. Thus it measures how well a classifier can discriminate between the outcome classes but does not take into account how well prediction scores are calibrated, i.e. prediction scores do not necessarily have to directly correspond to probabilities for the calculation of an AUC — they could be multiplied by any (identical) factor and still yield the same AUC.

If the outcome variable has more than two categories, we can calculate a weighted average of AUCs resulting from all combinations where one outcome category, the reference category, is seen as one class and all other observations form the other class [54, 138]. In the case of K classes, the average across K AUCs is calculated. Thus, in this one-versus-all approach one calculates

$$L_{\text{AUC}}(\hat{y}, y) = \sum_k^K L_{\text{AUC}, c_k}(\hat{y}, y) \cdot p(c_k)$$

with $L_{\text{AUC}}(\hat{y}, y)$ the AUC seen as loss function of the prediction \hat{y} and the true outcome y . L_{AUC, c_k} denotes the AUC based on class c_k and $p(c_k)$, the prevalence of class c_k in the data.

For evaluating if a classifier is calibrated well, other measures must be used in addition, for instance calibration plots or benefit curves [7].

2.4.2 Independent validation

When a prediction model has been trained it should be validated on an independent validation data set. Ideally one can train a final model on one cohort and validate it on another independent one.

If only one cohort is available and the number of given observations is limited cross-validation [76] is an efficient validation strategy even though it is, in fact, not a completely independent validation strategy. Note that if both, model selection and some generalizing performance of the best model, should be obtained within one cohort, nested cross-validation has to be applied (see Laimighofer et al. [106], for instance). Cross-validation often leads to high variance but this issue can be solved when special repeated cross-validation designs are used [61]. In the projects of the thesis external validation was possible, except for the project of Chapter 5 where the number of observations was that small that leave-one-out cross-validation was necessary to guarantee feasibility of fitting the classification models. Given the learning set size is fixed to $n - 1$, leave-one-out cross-validation is the most efficient validation resampling procedure regarding its variance [61].

Independent validation is also important for selecting the right model; for instance, for a LASSO model an appropriate tuning parameter λ has to be chosen with care. This tuning parameter or complexity parameter of a prediction model can be optimized in the scope of the so-called *bias-variance trade-off* [76]: Generally, the expectation (with respect to true outcome y and predictions \hat{y}) of the loss $L(y, \hat{y})$ can be decomposed into a noise term and variance and bias of \hat{y} . The noise term cannot be controlled, but variance and bias terms can be minimized; in parameter tuning, if the estimated model contains too many degrees of freedom, i.e. the free parameters that can vary independently, the model may overfit and have high variance when applied on independent data. On the other side, if the model does not contain a sufficient number of degrees of freedom, the model underfits the data and is highly biased. Both extrema usually lead to poor performance on independent data and a solution in between has to be found, for example by optimizing the complexity parameter via cross-validation. Figure 2.3 illustrates the situation for the LASSO.

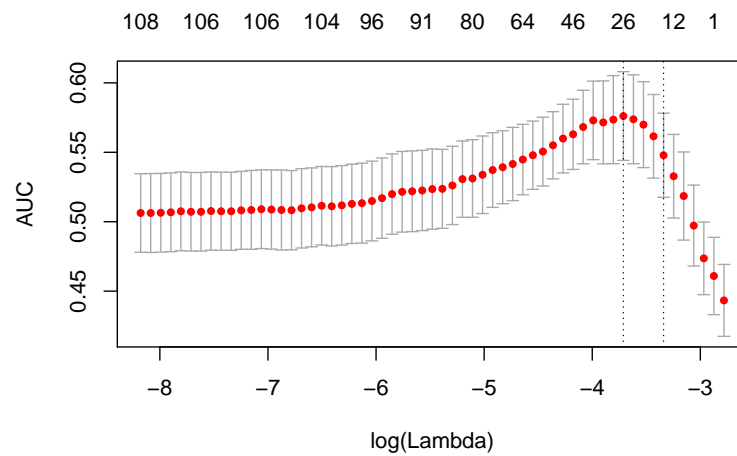


Figure 2.3: Test AUCs for optimizing tuning parameter λ via cross-validation in a LASSO model on toy data. The outcome variable is binary and predicted by up to $p = 110$ variables and $n = 500$ observations. The figure illustrates the bias-variance trade-off in model selection: Small and large values of λ lead to low AUCs and thus to poor performance. The optimal value for λ lies in between; the highest AUC is obtained when 26 of the 110 variables are selected, i.e. when 26 non-zero regression coefficients are estimated.

Chapter 3

Correcting classifiers for sample selection bias in two-phase case-control studies

Statistics is an art of inferring information about large populations from comparably small random samples. This is necessary because in practice it is most often impossible to receive measurements from all individuals in a population, e.g. due to organizational or cost reasons. In the clinical context, for example, one might aim to predict the risk for a certain disease based on clinical features for an entire population. The risk model will be derived from information from a much smaller random subsample of the population. When building such models, a common assumption is that the subsample follows the same distribution as the population the sample was taken from. This assumption, however, is not valid if the sample is not taken at random. In the epidemiological context, for example, this case occurs in the well-known *case-control studies* [154]. Here, one is interested in finding associations between features and rare disease outcomes. In order to increase precision and achieve higher statistical power for finding significant associations, cases are enriched such that cases and controls are equally represented in the sample. When a case-control study is used for risk prediction on an unbiased population e.g. via logistic regression, certain adjustments have to be made which have been elaborated in [84, 90, 146, 166].

An even more complex sample design appears in *two-phase case-control studies* [153, 183]. Here, one does not only enrich a rare disease outcome but also a rare covariate, e.g. an exposure. This measure prevents the sample from containing only few individuals which fall into both rare categories. From such a sample one would hardly be able to

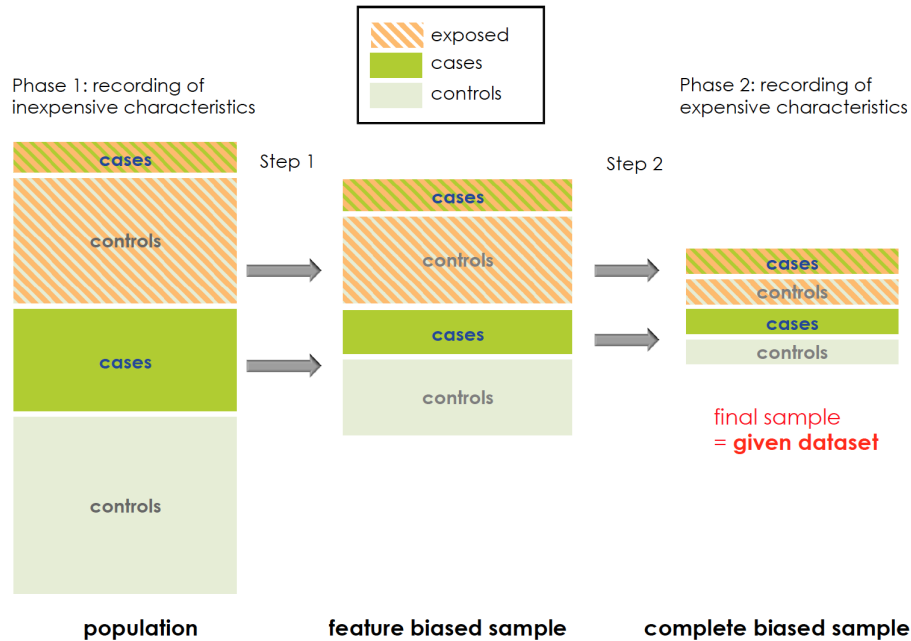


Figure 3.1a: Stratified random selection process of a two-phase case-control study. Feature characteristics known about a whole finite population are typically features which are inexpensive to measure and called characteristics recorded in Phase 1. The expensive characteristics are recorded only in Phase 2 — in the final sample.

draw conclusions about the rare combination. Figure 3.1a illustrates how the sampling procedure is performed in practice. Figure 3.1b shows an exemplary table of numbers of cases/controls and exposed/non-exposed individuals in the population and the sample. This and other complex survey designs, e.g. cohort sampling designs [150], have been used in order to obtain subpopulations with rare characteristics of features of interest [95, 121, 151]. The efficiency and analysis of the design are described in White [183].

In the situations described above the sample follows a different distribution than the population. This can affect statistical analysis. In the general context, the issue is known as *sample selection bias* [38, 77, 194]. It generally occurs when not all individuals from the population have the same probability of getting selected for the sample. If a statistical estimate is affected by sample selection bias, one should correct for it. The question of whether correction is necessary depends on the type of sample selection bias, the considered classifier, and the research question to be answered. For example, no adjustment is required if only the outcome variable is enriched and logistic regression is applied for prediction purposes, because the slope coefficients of the linear predictor remain asymptotically unaffected by sample selection bias for this case (if the functional form and the

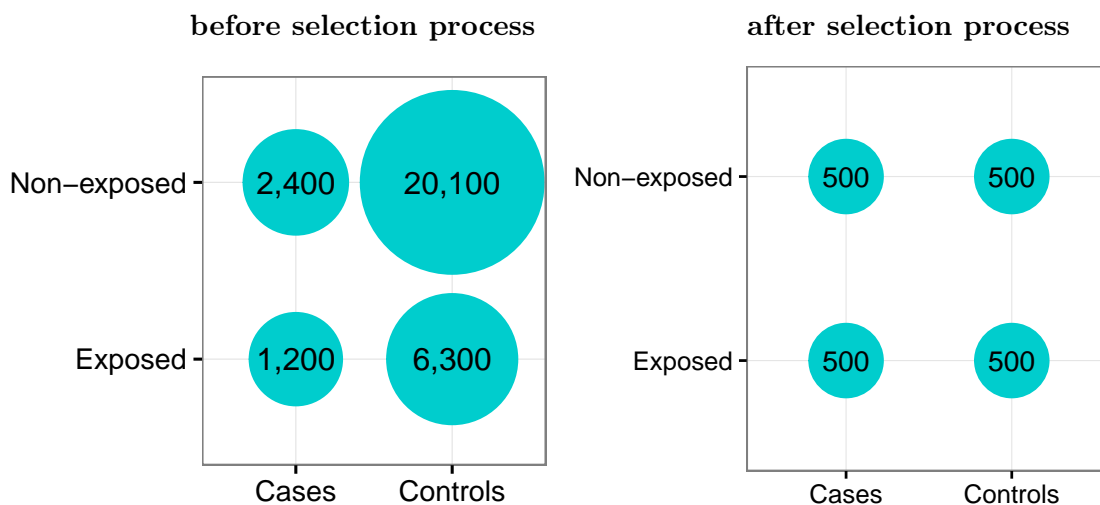


Figure 3.1b: Exemplary cross table for data before (left) and after (right) the selection process of a two-phase case-control study. There is a clear dependency between exposure and disease in the population. After the sampling process, this dependency vanishes completely for the final sample.

explanatory features for the model are correct) [97]. In general, however, correction is required, and there are several solutions to encounter this problem in complex survey designs [113, 179]. These existing approaches mainly focus on classical prediction methods or simple survey designs. Strategies applicable also for machine learning approaches have been suggested in the general sample selection bias context [52, 193, 194]. These methods reconstruct the population data or its covariance structure and typically involve non-parametric resampling techniques like bootstrapping. However, they neglect complex survey designs. Thus, while correcting for sample selection bias in logistic regression is well-investigated, its consideration is unclear for most machine learning approaches.

This chapter assesses, proposes and compares approaches to correct for sample selection bias in complex surveys, especially in two-phase case-control studies. Therefore we focus on binary outcome. Figure 3.2 illustrates the issue to be addressed. The emphasis is on a widely-used machine learning approach: the random forest. We correct for the covariance structure of the sample by incorporating knowledge about the sample selection procedure into nonparametric and parametric resampling techniques. As the random forest is based on resampling anyway (in terms of bagging, see Section 3.2.4), we incorporate the correction step into the inherent resampling procedure. We compare our correction approaches to analogous state-of-the-art approaches, both for the random forest and other common classifiers, namely logistic regression, logistic regression including interaction terms, and the naive Bayes classifier. We especially address the question whether correction is neces-

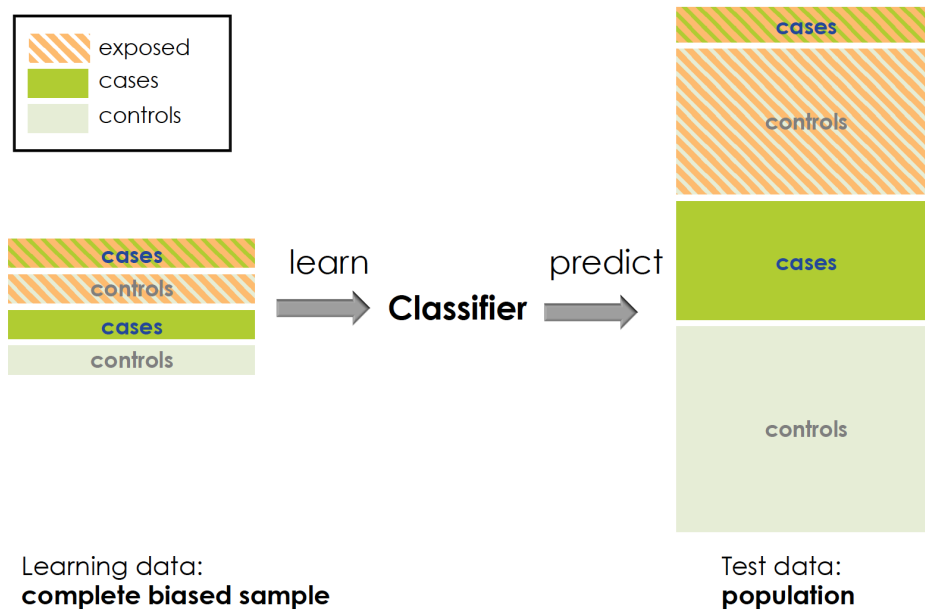


Figure 3.2: Scheme of learning on biased learning data and predicting on unbiased test data. The classifier learns on four equally sized strata (complete biased learning data set) but predicts on a data set (unbiased population) of different sizes of the four strata.

sary in random forests, and if so, whether current correction approaches can successfully be transferred to the random forest and whether improvement is possible through alternative approaches. We assess and compare the prediction performance of the correction techniques in a synthetic simulation study and in a real data application. We provide the R package *sambia* so that readers can easily apply the methods presented here to their data.

This chapter is structured as follows: We formalize sample selection bias and address the necessity of correction in Section 3. Section 3.2 explains current approaches for corrected learning on biased samples, and we propose two new methods based on drawing observations from theoretical distributions assumed for the given data. We furthermore analyze properties of the various approaches in the context of sample selection bias. Section 3.3 presents a simulation study which compares all approaches regarding performance on new unbiased test data. Section 3.4 shows a similar analysis on real data. We discuss and conclude our work in Section 3.5.

This chapter is in parts identical with the following publication:

[100]: Norbert Krautenbacher, Fabian J. Theis, and Christiane Fuchs. Correcting Clas-

sifiers for Sample Selection Bias in Two-Phase Case-Control Studies. *Computational and Mathematical Methods in Medicine*, 2017:18, 2017. doi: 10.1155/2017/7847531

3.1 Sample selection bias and stratified random sampling

This section introduces general definitions and background information: A formal description of sample selection bias (Section 3.1.1), the special case of two-phase case-control studies (Section 3.1.2), and properties of biased samples (Section 3.1.3).

3.1.1 Sample selection bias — definition

The following set-up is similar to Zadrozny [194] and distinguishes *sample selection bias* into three types. Throughout this chapter, \mathcal{Y} is a discrete *binary* label space since we focus on binary classifiers in this chapter. General notation has been described in Section 2.1.

For the set-up of the sample selection bias issue, let \mathcal{S} be a binary space. $S \in \mathcal{S}$ is the variable that controls the selection of observations: For $s_i = 1$ the i th observation is selected, for $s_i = 0$ the observation is not selected. Thus, observations (\mathbf{x}_i, y_i, s_i) are drawn from a distribution \mathcal{L} with domain $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$.

In general, a sample $\{(\mathbf{x}_i, y_i, s_i)\}_{i=1, \dots, n}$ can be biased in three different ways. These types of sample selection bias can be described as follows [52, 194] :

- *Label bias*: biasedness depends on Y only, so $P(S|\mathbf{X}, Y) = P(S|Y)$ but $P(S|Y) \neq P(S)$.
- *Feature Bias*: biasedness depends on \mathbf{X} only, so $P(S|\mathbf{X}, Y) = P(S|\mathbf{X})$ but $P(S|\mathbf{X}) \neq P(S)$.
- *Complete Bias*: biasedness depends on \mathbf{X} and Y , i. e. there is no independence between S and \mathbf{X}, Y , so $P(S|\mathbf{X}, Y) \neq P(S|Y)$ and $P(S|\mathbf{X}, Y) \neq P(S|\mathbf{X})$.

Under label bias, S is not necessarily independent of \mathbf{X} [52], and for feature bias S is not necessarily independent of Y . This follows from

$$P(S|\mathbf{X}, Y) = P(S|Y) \wedge P(S|Y) \neq P(S) \not\Rightarrow P(S|\mathbf{X}) = P(S). \quad (3.1)$$

This holds since: if $\mathbf{x} := t(y)$ where t is a function mapping to $\{0, 1\}$, then $P(S|\mathbf{x}) = P(S|t(y)) = P(S|y) \neq P(S)$.

Analogously, one can show that feature bias does not imply that S is independent of Y .

Whenever there is sample selection bias, there are *selection probabilities* $P(S = 1|Y, \mathbf{X})$ (in particular $P(S = 1|Y)$ for label bias and $P(S = 1|\mathbf{X})$ for feature bias). In practice, these probabilities can often be estimated if they are unknown (see Huang et al. [83], for instance). However, when sample selection bias has arisen due to a certain sampling design they are typically known. Since throughout this chapter this case is treated exclusively it is reasonable to assume them to be provided. All approaches proposed in this chapter will incorporate these selection probabilities in terms of weights corresponding to the inverse probabilities $P(S = 1|\mathbf{X}, Y)^{-1}$.

3.1.2 Two-phase case-control studies

In this chapter we will treat the special case of two-phase case-control studies and hence put them into the context of sample selection bias in this subsection.

The case-control study is an example for sample selection bias in the clinical context: Some diseases under investigation are very rare in the entire population. A random sample of study participants would contain very few cases of the disease. Statistical analysis would suffer from low precision and thus low power. In order to increase precision and power, the number of cases is enriched such that the proportion of cases and controls in the sample is identical. In particular, $P(Y = 1|S) = 0.5$ whereas the prevalence rate $P(Y = 1)$ is much smaller, so $P(Y = 1|S) \neq P(Y = 1)$. This by Bayes' theorem implies $P(S|Y = 1) \neq P(S)$, and thus there is label bias.

Case-control studies are mostly used for investigating associations between disease and features. The underlying label bias does not alter the effect estimates in hypothesis testing for associations between disease and features. However, this is true only asymptotically,

and there may be consequences in small sample scenarios. If one focuses on prediction e. g. via logistic regression — as we do in this chapter — the intercept estimate can simply be adjusted as described in Rose and van der Laan [146] or Steyerberg et al. [166]. Elkan [50] offers a solution for arbitrary classifiers.

In *two-phase case-control studies*, on the other hand, the selection is additionally controlled by a categorical feature variable. Such studies suffer from label *and* feature bias, so there is complete bias. We focus on this case, i. e. complex survey designs which involve complete bias.

3.1.3 Stratified random samples

When data is sampled as in one-phase or two-phase case-control studies, there are groups within which the selection probabilities are equal. These groups are called *strata*. In this chapter we focus on two-phase case-control studies where the strata are determined by a categorical stratum feature (often an exposure) X_e and the outcome Y . The remaining features of \mathbf{X} are $\tilde{\mathbf{X}} := \mathbf{X} \setminus X_e$.

For a population of size N and sample size n let $h \in \{1, \dots, H\}$ be the index of the stratum. Realizations falling into stratum h are denoted by $\tilde{\mathbf{x}}_h$, x_{eh} and y_h , or combined as $(\mathbf{x}_h, y_h) = (\tilde{\mathbf{x}}_h, x_{eh}, y_h)$. We denote by n_h the size of the stratum h in the sample and by N_h its size in the population. Then clearly $P(S = 1) = \frac{n}{N}$ and

$$P(S = 1|\mathbf{x}, y) = P(S = 1|x_e, y) = P(S = 1|h(x_e, y)) = \frac{n_{h(x_e, y)}}{N_{h(x_e, y)}}, \quad (3.2)$$

where $h(x_e, y)$ denotes the stratum determined by x_e and y . Throughout the chapter we will simply abbreviate this by h .

If the features determining the selection probabilities are categorical, the data set can be partitioned into corresponding strata with equal selection probabilities. This is not the case if e. g. the feature causing the selection bias is continuous. In the categorical case, selection probabilities can be used for adjusting the distribution of the sample to the original distribution of the population.

Consider the selection probability $P(S = 1|h)$ for an observation of stratum h . We define

$$w_h := \left\lceil \frac{\max_{h'} P(S = 1|h')}{P(S = 1|h)} \right\rceil$$

as the *inverse-probability (IP) weight* for stratum h . The squared brackets denote rounding

to the closest integer. The term *IP weight* is sometimes used in literature for the simple inverse selection probability $P(S = 1|h)^{-1}$. In this work, we use w_h rather than $P(S = 1|h)^{-1}$ to keep the number of newly generated observations minimal.

In our correction approaches we will use

$$n' := \sum_{h=1}^H n_h w_h, \quad (3.3)$$

which can be seen as the number of re-weighted observations, i. e. the sum of all observations multiplied by their weights. As stated above we are interested in adjustment methods which can be applied to arbitrary classifiers. In the next section, after stating a typical set-up of a statistical learning procedure, we will describe several sample selection bias correction approaches proposed in literature.

3.2 Methods

In this section we describe, modify and analyze IP weight-incorporating classifiers which are designed for learning on an unbiased data set, when only a biased data set for learning is given.

All correction approaches adjust the given data set to correct for sample selection bias by reconstructing the original (unbiased) data structure before or while learning the classifier. As introduced in Chapter 2 we consider the classifier

$$\varphi : \begin{cases} (\mathcal{X} \times \mathcal{Y})^{\times n} \times \mathcal{X} & \rightarrow \mathcal{Y} \\ ((\mathbf{x}, \mathbf{y}), \mathbf{X}) & \mapsto \varphi((\mathbf{x}, \mathbf{y}); \mathbf{X}), \end{cases}$$

where the given learning data set $(\mathbf{x}, \mathbf{y}) = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n))$ is mapped to the prediction (in our case classification) rule and applied to the random variable \mathbf{X} .

3.2.1 State of the art correction approaches

The methods in this section were proposed in literature and are partly modified for our purposes.

No correction

The naive approach for learning on a biased sample is to simply ignore the bias. No IP weights are used, and the classifier is trained on the given sample as it is. As shown in Zadrozny [194], this approach is valid for some cases of sample selection bias, namely for feature bias for a specific type of classifiers.

Inverse-probability oversampling

An intuitive method for correcting for sample selection bias is the plain replication of each observation in the sample according to its IP weight, i.e. in a stratified random sample one replicates an observation of stratum h by the factor w_h . Then, the number of observations in the reconstructed sample is n' . This sample is used for learning. In maximum likelihood-based approaches like generalized regression models, this method is equal to weighting the single likelihoods per observation. The procedure, sometimes simply called *inverse-probability weighting*, has been used early [82], with applications both in regression [144] and general statistical learning [50]. We refer to this technique as *IP oversampling*: since in the stratification process some observations were *oversampled*, this method is a way of re-oversampling under-represented observations in the stratified sample. Since IP oversampling is applicable to arbitrary classifiers, we take it into account for further comparisons. A drawback is that it changes the covariance structure per stratum h . In Section 3.2.2 we propose a method that corrects for this issue.

Inverse-probability bagging

Another correction method uses bootstrap aggregation and averaging, commonly abbreviated to the acronym *bagging*. The procedure averages several predictions trained on an ensemble of bootstrap samples and thus makes learners more robust [25]. Nonparametric bootstrap samples arise by randomly drawing n times from the original data set of size n with replacement. Bagging procedures fit a learner on each of these bootstrap samples and combine the learners by averaging predictions or by majority vote. When building

bootstrap samples from biased data sets, as in our case, resampling can take into account IP weights: instead of drawing observations randomly, selection probabilities are set proportional to w_h for the respective strata h . This procedure is proposed in Nahorniak et al. [126] and labeled *IP bagging* here.

Costing

Zadrozny et al. [193] argue that sampling with replacement as done in IP bagging is inappropriate since sets of independent observations from continuous distributions contain two identical elements only with zero probability, whereas nonparametric bootstrap samples generally contain observations repeatedly. Zadrozny et al. [193] propose an approach called *costing*, which is similar to IP bagging in terms of resampling from the learning data and aggregation of learned algorithms on m new samples. It differs in the implementation of resampling the m learning sets: here, an observation from the original learning set enters a resampled data set only once at most. It is selected with probability $w_h / \max_{h'} w_{h'}$ according to the corresponding stratum h . Consequently the size of the new samples is smaller than n and generally varies among the m learning sets. The latter aspect indicates the difference of this approach to subsampling without replacement. A detailed description of the aspects of the algorithm can be found in Zadrozny et al. [193], Sections 2.3.2 to 2.3.4.

A drawback of costing in case of strata with a low number of observations is the following: there may be subsamples which do not contain observations from all strata, which implies that no classification rule can be learned for the missing strata from those subsamples. For the purposes of this project, we adjusted the costing algorithm by not taking into account such incomplete samples. This modification causes bias which we consider negligible.

Modified SMOTE

So far, all correction approaches replicated given observations. In contrast, Chawla et al. [34] propose a *synthetic minority over-sampling technique (SMOTE)* to generate new, synthetic data. The strategy is designed as a solution for the imbalanced class problem, where rare cases (the *minority class*) are hardly represented in the (non-stratified) sample, which mainly consist of common cases from the *majority class*. In this situation, several

classifiers perform poorly because of the imbalanced proportion of outcome categories in the data.

In its original form, SMOTE generates synthetic observations for the minority class as follows: For fixed $k \in \mathbb{N}$, one determines the k nearest neighbors of the minority class. Depending on the desired number of new observations, one then randomly selects an according amount of instances from this neighborhood. New observations arise as weighted averages between original feature vectors and selected nearest neighbors. To that end, weights are randomly sampled from the unit interval.

We adapt SMOTE to the context of stratified random samples: rather than enlarging only the minority class, we generate synthetic observations for all strata with $w_h > 1$. Thus we apply SMOTE up to $H - 1$ times, once for each stratum which requires more observations. We refer to this algorithm as *modified SMOTE* hereafter.

3.2.2 Correcting covariance structures

The approaches above aim to reconstruct the original data distribution in order to then learn a classifier on an unbiased sample. However, several aspects are not incorporated so far: IP oversampling replicates observations and by this biases the covariance-structure within the strata. A correction for this biasedness should be provided. Similarly, modified SMOTE biases the data, especially for large weights w_h , where the same observations are used several times for synthetic data generation and lack contributing sufficient variation. IP bagging and costing both exclusively base on resampling observed data. This may become problematic especially for small sample sizes or only small stratum sizes (which can occur in the resampled data sets for these two approaches): the fine structure in the given data can be spurious due to the deficit of observations. Also due to small sample sizes and hence too few values in the sample only covering a restricted range one may underestimate variance and covariance of the data.

In this section we propose two procedures which aim to conquer the problem of small strata by increasing the number of observations per stratum and at the same time estimate the covariance of the population appropriately. The idea behind both approaches is to exploit the fact that within each stratum h all observations are assigned the same weight w_h . This enables parametric resampling within each stratum.

Let $\tilde{\mathcal{L}}_h$ be the distribution which $\tilde{\mathbf{X}}_h$ follows. We aim to approximate $\tilde{\mathcal{L}}_h$ by theoretical distributions and estimate their parameters for each stratum h . In practice, determining

the multivariate distribution of the features is difficult and relies on assumptions. One might e. g. assume normally distributed features,

$$\tilde{\mathbf{X}}_h \sim \mathcal{N}(\mu_h, \Sigma_h), \quad (3.4)$$

and would then have to estimate $\hat{\mu}_h$ and $\hat{\Sigma}_h$ for all h , which is typically done by their empirical pendants. Even though we focus on the normal distribution in our empirical investigations, we propose the following approaches such that they can be applied to arbitrary distribution assumptions.

Stochastic inverse-probability oversampling

Our first approach builds upon the re- or oversampling techniques described in Section 3.2.1. However, the repeated occurrence of observations of continuous features falsifies the covariance structure of the reconstructed samples. Hence, we add noise to those data sets obtained via IP oversampling and thus call our proceeding *stochastic IP oversampling*.

When adding this noise, we want to retain important distribution characteristics of the respective stratum. As stated above, the stratified sample contains features $\tilde{\mathbf{X}}_h \sim \tilde{\mathcal{L}}_h$. After performing IP oversampling, the reconstructed features $\tilde{\mathbf{X}}'_h$ do not follow $\tilde{\mathcal{L}}_h$ anymore. We aim to adjust $\tilde{\mathbf{X}}'_h$ by adding noise terms $\tilde{\boldsymbol{\varepsilon}}_h$ such that $\tilde{\mathbf{X}}'_h + \tilde{\boldsymbol{\varepsilon}}_h$ approximately follows the original distribution $\tilde{\mathcal{L}}_h$ in the sense that it agrees in expectation and covariance. In the following we derive a respective distribution $\tilde{\mathcal{L}}_h^{adj}$ for $\tilde{\boldsymbol{\varepsilon}}_h$.

We seek two conditions to hold:

$$\mathbb{E}(\tilde{\mathbf{X}}'_h + \tilde{\boldsymbol{\varepsilon}}_h) = \mathbb{E}(\tilde{\mathbf{X}}_h) \quad (3.5)$$

$$\text{Cov}(\tilde{X}_h^{(k)'} + \tilde{\varepsilon}_h^{(k)}, \tilde{X}_h^{(j)'} + \tilde{\varepsilon}_h^{(j)}) = \text{Cov}(\tilde{X}_h^{(k)}, \tilde{X}_h^{(j)}) = \Sigma_h \quad (3.6)$$

for all $k, j \in \{1, \dots, p\}$ here denoting the index of the features. Because of (3.5) and since $\mathbb{E}(\tilde{\mathbf{X}}'_h) = \mathbb{E}(\tilde{\mathbf{X}}_h)$ we obtain

$$\mathbb{E}(\tilde{\boldsymbol{\varepsilon}}_h) = 0. \quad (3.7)$$

The adjusted *noise covariance matrix* $\Sigma_h^{adj} := \text{Cov}(\tilde{\varepsilon}_h^{(k)}, \tilde{\varepsilon}_h^{(j)})$ can be written as

$$\Sigma_h^{adj} = \frac{w_h - 1}{w_h n_h - 1} \Sigma_h. \quad (3.8)$$

Proof of Equation 3.8: We derive an appropriate noise covariance matrix to be added to the features $\tilde{\mathbf{X}}_h'$ resulting from IP oversampling: For one stratum h we look at the covariance of the pair of features $\tilde{X}_h^{(k)}, \tilde{X}_h^{(j)}$ for $k, j \in \{1, \dots, p\}$. For sample size n , we get per stratum a sample covariance per pair $\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}$, given by

$$s_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\tilde{x}_{hi}^{(k)} - \bar{x}_h^{(k)}) (\tilde{x}_{hi}^{(j)} - \bar{x}_h^{(j)}),$$

where $\bar{x}_h^{(l)} := \frac{1}{n_h} \sum_{i=1}^{n_h} \tilde{x}_{hi}^{(l)}$ for any $l \in \{1, \dots, p\}$.

For IP oversampling we replicate the data points by the factor w_h , which varies per stratum. Thus the covariance of the modified sample is

$$\begin{aligned} s'_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}} &= \frac{1}{w_h n_h - 1} \sum_{i=1}^{n_h} w_h (\tilde{x}_{hi}^{(k)} - \bar{x}_h^{(k)}) (\tilde{x}_{hi}^{(j)} - \bar{x}_h^{(j)}) \\ &= \frac{w_h (n_h - 1)}{w_h n_h - 1} s_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}}. \end{aligned} \quad (3.9)$$

In addition to simple IP oversampling, stochastic IP oversampling incorporates the summation of some noise (matrix) $\tilde{\varepsilon}$. We want the following to hold for a pair of the random vectors $\tilde{\varepsilon}^{(k)}, \tilde{\varepsilon}^{(j)}$ of size n_h :

$$\text{Cov}(\tilde{X}_h^{(k)'} + \tilde{\varepsilon}_h^{(k)}, \tilde{X}_h^{(j)'} + \tilde{\varepsilon}_h^{(j)}) = \text{Cov}(\tilde{X}_h^{(k)}, \tilde{X}_h^{(j)}), \quad (3.10)$$

where $\tilde{X}_h^{(k)'}, \tilde{X}_h^{(j)'}$ are the random variables resulting from replication by a factor w_h (oversampling).

We can simplify

$$\begin{aligned} \text{Cov}(\tilde{X}_h^{(k)'} + \tilde{\varepsilon}_h^{(k)}, \tilde{X}_h^{(j)'} + \tilde{\varepsilon}_h^{(j)}) &= \text{Cov}(\tilde{X}_h^{(k)'}, \tilde{X}_h^{(j)'}) + \text{Cov}(\tilde{X}_h^{(k)'}, \tilde{\varepsilon}_h^{(j)}) \\ &\quad + \text{Cov}(\tilde{X}_h^{(j)'}, \tilde{\varepsilon}_h^{(k)}) + \text{Cov}(\tilde{\varepsilon}_h^{(k)}, \tilde{\varepsilon}_h^{(j)}) \\ &= \text{Cov}(\tilde{X}_h^{(k)'}, \tilde{X}_h^{(j)'}) + \text{Cov}(\tilde{\varepsilon}_h^{(k)}, \tilde{\varepsilon}_h^{(j)}), \end{aligned}$$

since the noise component $\tilde{\varepsilon}_h^{(j)}$ should not correlate with the feature random vector \mathbf{X}_k (neither $\tilde{\varepsilon}_h^{(k)}$ with $\tilde{X}_h^{(j)}$, respectively). This also holds for $j = k$.

We can estimate the components of the covariance matrix $\text{Cov}(\tilde{X}_h^{(k)'}, \tilde{X}_h^{(j)'})$ by $s'_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}} =$

$\frac{w_h(n_h-1)}{w_h n_h - 1} s_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}}$. Substituting this into (3.10) yields for the entries of our *noise covariance matrix*:

$$s'_{\tilde{\varepsilon}_h^{(k)}, \tilde{\varepsilon}_h^{(j)}} = s_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}} - \frac{w_h(n_h-1)}{w_h n_h - 1} s_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}} = \frac{w_h-1}{w_h n_h - 1} s_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}}.$$

In terms of random variables, the empirical covariance matrix combining all entries $s'_{\tilde{\varepsilon}_h^{(k)}, \tilde{\varepsilon}_h^{(j)}}$ for all $k, j \in \{1, \dots, p\}$ would be replaced by Σ_h^{adj} and the empirical covariance matrix combining all entries $s_{\tilde{x}_h^{(k)}, \tilde{x}_h^{(j)}}$ for all $k, j \in \{1, \dots, p\}$ by Σ_h . \square

For instance, when assuming a multivariate normal distribution $\tilde{\mathbf{X}}_h \sim \tilde{\mathcal{L}}_h = \mathcal{N}(\mu_h, \Sigma_h)$, the noise term

$$\tilde{\varepsilon}_h \sim \tilde{\mathcal{L}}_h^{adj} = \mathcal{N}\left(0, \frac{w_h-1}{w_h n_h - 1} \Sigma_h\right) \quad (3.11)$$

would retain the stratum expectation and covariance (and thus in the Gaussian case the entire distribution).

In order to make a corresponding correction method more robust, we repeat the noise-adding procedure and average over the models fitted on each of those repetitions. Algorithm 3.1 displays the single steps of stochastic IP oversampling.

Parametric inverse-probability bagging

Stochastic IP oversampling above consisted of a deterministic replication of observations followed by a stochastic alteration by adding noise. Now, we propose a completely parametric approach which we call *parametric IP bagging*. As in IP bagging, we draw bootstrap samples from the original stratified data set. This time, however, we employ parametric instead of non-parametric bootstrap and set the bootstrap sample size to n' . As in stochastic IP oversampling, we assume a multivariate distribution underlying the original data and estimate the parameters stratum-wise. The procedure is defined by Algorithm 3.2.

Algorithm 3.1: Stochastic inverse-probability oversampling

Input: Observed sample $(\tilde{\mathbf{x}}, x_e, y)$ of size n , IP weights w_h **Output:** Unbiased prediction \hat{y} for new unbiased data $(\mathbf{X}, Y) \sim D$ 1. Perform IP oversampling, resulting in reconstructed sample $(\tilde{\mathbf{x}}', x_e', y')$ of size n' 2. **for** $b = 1$ to B **do** **for** $h = 1$ to H **do** (a) Estimate Σ_h^{adj} of distribution $\tilde{\mathcal{L}}_h$ (b) Draw noise vector $\tilde{\boldsymbol{\varepsilon}}_h^b$ from $\hat{\mathcal{L}}_h^{adj}$ of length $n_h w_h$ (c) Rebuild original stratum as $(\tilde{\mathbf{x}}'_h + \tilde{\boldsymbol{\varepsilon}}_h^b, x_{e'_h}, y'_h)$ **end**

(a) Combine strata to sample:

$$(\tilde{\mathbf{x}}' + \tilde{\boldsymbol{\varepsilon}}^b, x_{e'}', y') = ((\tilde{\mathbf{x}}'_1 + \tilde{\boldsymbol{\varepsilon}}_1^b, x_{e'_1}, y'_1), \dots, (\tilde{\mathbf{x}}'_H + \tilde{\boldsymbol{\varepsilon}}_H^b, x_{e'_H}, y'_H))$$

 (b) Fit classifier $\hat{y}^b = \varphi((\tilde{\mathbf{x}}' + \tilde{\boldsymbol{\varepsilon}}^b, x_{e'}', y'); \mathbf{X})$ **end**3. Output the ensemble of learners $\{\hat{y}^b\}_{b=1, \dots, B}$ 4. Aggregate predictions on new data set by averaging: $\hat{y} = \sum_{b=1}^B \hat{y}^b$

3.2.3 Properties of correction approaches

So far, we described seven ways to deal with sample selection bias: no correction, IP oversampling, IP bagging, costing, modified SMOTE, stochastic IP oversampling and parametric IP bagging. This subsection compares their characteristics. They are summarized in the left part of Table of 3.1 on page 55.

(i) Incorporation of weights

Except for the non-correction approach, all correction methods incorporate weights. As mentioned in 3.2.1 there are cases of sample selection bias where the bias does not affect the classifier so that correction in terms of weighting is not necessary. However, as we will elaborate in this chapter on two-phase case-control studies, correction is necessary in the context of complete bias.

Algorithm 3.2: Parametric inverse-probability bagging

Input: Observed sample $(\tilde{\mathbf{x}}, x_e, y)$ of size n , IP weights w_h **Output:** Unbiased prediction \hat{y} for new unbiased data $(\mathbf{X}, Y) \sim D$

1. **for** $b = 1$ to B **do**
 - for** $h = 1$ to H **do**
 - (a) Estimate parameters of distribution $\tilde{\mathcal{L}}_h$
 - (b) Draw parametric bootstrap sample $\tilde{\mathbf{x}}_h^b$ from $\tilde{\mathcal{L}}_h$ of size $n_h w_h$
 - (c) Rebuild stratum as $(\tilde{\mathbf{x}}_h^b, x_e^{\times w_h}, y_h^{\times w_h})$, where ' $\times w_h$ ' denotes w_h -fold concatenation
 - end**
 - (a) Combine strata to sample:
 $(\tilde{\mathbf{x}}^b, x_e^{\times w}, y^{\times w}) = ((\tilde{\mathbf{x}}_1^b, x_{e1}^{\times w_1}, y_1^{\times w_1}), \dots, (\tilde{\mathbf{x}}_H^b, x_{eH}^{\times w_H}, y_H^{\times w_H}))$
with $w = \sum_{h=1}^H w_h$
 - (b) Fit classifier $\hat{y}^b = \varphi((\tilde{\mathbf{x}}^b, x_e^{\times w}, y^{\times w}); \mathbf{X})$
 - end**
 2. Output the ensemble of learners $\{\hat{y}^b\}_{b=1, \dots, B}$
 3. Aggregate predictions on new data set by averaging: $\hat{y} = \sum_{b=1}^B \hat{y}^b$
-

(ii) Correcting covariance structure of learning data

Sample selection bias can cause a biased covariance structure in the data. Some but not all correction approaches correct for this bias: The non-correction approach clearly uses the biased covariance structure. Also IP oversampling does not correct for it; the replication of observations generally leads to underestimating the covariance (cf. Equation 3.9). For modified SMOTE, the resulting covariance structure depends on the magnitude of the weights w_h and the degree of separation of the features into distinct clusters. For instance, a stratum with large weight w_h will cause a large number of newly generated observations as compared to the original number observations. The same neighbours will be selected several times such that sufficient variation of the new observations cannot be guaranteed. This may result in a similar issue as for IP oversampling described above. All other approaches aim to obtain the right covariance structure per stratum and in the entire reconstructed sample.

(iii) Size of reconstructed samples

As a well-known fact in statistical learning, the bias of a classifier increases when the learning sample size decreases. IP bagging is based on reconstructed samples of the same size n as the original stratified data set. Sample sizes in costing are even smaller and vary between bootstrap samples. Especially the small strata contain a small number of observations for these two ways of reconstructing the sample. Consequently a certain structure of the data may get lost for learning, e.g. the appropriate variability within small strata may not be given anymore. IP oversampling, modified SMOTE and our own methods stochastic IP oversampling and parametric IP bagging, on the other hand, employ reconstructed samples of larger sizes n' as defined in Equation (3.3). By this we intend to have sufficient numbers of observations in each stratum for possibly improving the learning of the classifier as compared to the use of smaller samples. In the non-parametric IP oversampling, the larger sample size induces a large number of perfectly repeated observations. This, again, biases the covariance structure. In our parametric approaches, stochastic IP oversampling and parametric IP bagging, this drawback does not occur.

3.2.4 Classifiers

In Sections 3.2.1 and 3.2.2 several approaches adjusting for sample selection bias have been presented and proposed. We implemented all approaches for the following classifiers: classical logistic regression based on maximum likelihood estimation as a classifier serving as reference since correction approaches are well-established for it; the tree-based random forest as our main object of interest; logistic regression including interaction terms; and the naive Bayes classifier as further algorithms for comparison. These classifiers have been described in Sections 2.3.1, 2.3.6, and 2.3.8.

For logistic regression we investigate two variants of this model: Once, all features enter the model just linearly. In a refinement, features are additionally included as all possible two-way interaction term combinations, not only in order to detect possible interaction effects but also to obtain more complex decision boundaries.

As described in Zadrozny [194], a classifiers' output can either depend on $P(Y|\mathbf{x})$ only or on both $P(Y|\mathbf{x})$ and $P(\mathbf{X})$. The first type of classifiers per definition is not affected by

feature bias whereas the second type is affected. Thus one has to consider that the two types behave differently under complete bias, as well.

We did the following adjustment in the random forest procedure: For all approaches in Section 3.2.1 and 3.2.2 which are based on aggregating after re-sampling, namely IP bagging, costing, stochastic IP oversampling, and parametric IP bagging, we incorporate these approaches into the random forest correspondingly. That means, instead of performing bagging within another bagging, we combine the two procedures. Note that IP oversampling incorporated in a random forest turns the approach to a bagging method. In fact, IP oversampling is exactly the same method as IP bagging when using samples of size n' instead of n . Thus for the implementation of our approaches into the random forest we implicitly take both versions of IP bagging into account.

3.3 Simulation study

So far, we have presented and developed strategies for fitting classifiers under complete bias. In this section we investigate their performance when a sample from a two-phase case-control study is given as learning data set but the test data is unbiased, i.e. it is a random sample from the population. We do this in a simulation study. After stating the set-up in Section 3.3.1, we compare performances for the introduced correction approaches and classifiers (Section 3.2) and report the results in Section 3.3.2.

3.3.1 Design

For evaluating the performance of correction approaches on training samples from two-phase case-control studies and unbiased validation data sets, we need three kinds of data sets: First, a biased learning data set stemming from a two-phase case-control study. Second, an unbiased large reference learning data set for comparison purposes. We refer to this data as *population*. It is not available in practice. Third, an unbiased test data set distributed like the population is required. We artificially simulated such data sets as described in the following.

We started by generating the large unbiased population data set. To that end, we randomly sampled 10^5 feature vectors consisting of one binary exposure variable X_e and $p = 5$ continuous other features $\tilde{X}^{(j)}$, $j \in \{1, \dots, 5\}$. The exposure X_e was meant to serve as a stratum feature with a low proportion (10%) of exposed ($X_e = 1$) individuals and a majority of non-exposed ($X_e = 0$) individuals. The $p = 5$ other features were generated independently of x_e and of each other. We investigated the following four distribution families:

- Normal distribution: $\tilde{X}^{(j)} \sim \mathcal{N}(\mu^{(j)}, \sigma^{(j)2})$ for all $j = 1, \dots, p$,
- Student's t distribution: $\tilde{X}^{(j)} \sim t(v_j)$ for all $j = 1, \dots, p$,
- Poisson distribution: $\tilde{X}^{(j)} \sim \text{Po}(\lambda_j)$ for all $j = 1, \dots, p$,
- Bernoulli distribution: $\tilde{X}^{(j)} \sim \text{Ber}(\pi_j)$ for all $j = 1, \dots, p$.

The distribution parameters were uniformly drawn from the following sets for $j = 1, \dots, p$: mean $\mu^{(j)} \in [1, 10]$, standard deviation $\sigma^{(j)} \in [1, 5]$, degrees of freedom $v_j \in \{10, 11, 12, \dots, 98, 99, 100\}$, event rate $\lambda_j \in \{1, 2, 3, 4, 5\}$, probability of success $\pi_j \in [0.4, 0.6]$.

In order to also investigate more realistic distribution scenarios, we additionally generated and analyzed data sets with dependent features and features from different distributions. These studies yield similar results as the setting above and are described in the supplementary material for this chapter (see Supplemental Section A.1.1).

Given the covariates $\mathbf{X} = (X_e, \tilde{\mathbf{X}})$, the outcome Y was generated according to a logistic regression model: $Y|\mathbf{X} \sim \text{Ber}(\theta(\mathbf{X}))$, where $\theta(\mathbf{X}) = (1 + \exp\{-(\beta_0 + \mathbf{X}\boldsymbol{\beta})\})^{-1}$. We chose the effects in terms of regression coefficients $\boldsymbol{\beta} = (\beta_e, \beta_1, \dots, \beta_5)'$ as follows: The exposure has a negative effect on the outcome with $\beta_e := \log 0.5$. The effects β_1, \dots, β_5 for the main features are varied at random, namely uniformly on the interval $[-0.15, 0.15]$ in order to gain an intermediate performance of a classifier applied on an independent data set. β_0 was chosen such that $P(Y = 1) = 0.1$. By this setup the population with a rare exposure, $P(X_e = 1) = 0.1$, and rare cases, $P(Y = 1) = 0.1$, is fully generated. In the Supplemental Section A.1.2 we treat further simulations scenarios where these probabilities and the sample sizes are varied.

In order to obtain a biased stratified sample, we simulated a two-phase random selection process from the population (Figure 3.1a) such that $P(Y = 1|S) = 0.5$ **and** $P(X_e = 1|S) = 0.5$. In a first step an equal number of observations was randomly taken with $x_e = 1$ and with $x_e = 0$. In a second step, in each of these two strata from the first

step, an equal number of observations with $y = 1$ and $y = 0$ was selected. By this we partitioned the population into four equally-sized strata corresponding to $(y, x_e) \in \{(1, 1), (1, 0), (0, 1), (0, 0)\}$. By Supplemental Section A.1.3 we also cover the scenarios for one-phase random selection processes, i.e. when there is only label bias or only feature bias.

Test data sets of size 10^4 were created in exactly the same way as the population. For our simulation study, we generated the population data set, the stratified data set and the test set 1000 times for each feature distribution assumption. This way, we could empirically assess the variability of the performance of the correction and classification methods.

Looking at the population's univariate marginal distributions of the single variables and comparing them to those of a biased sample arisen as described above, and to corrected samples which were generated by parametric IP bootstrap indicates that the distribution in the population is more similar to the corrected sample distribution than it is to the distribution of the biased sample. (see Figure 3.3).

Application of classifiers

We apply the seven correction approaches (Section 3.2) combined with the four considered classifiers (Section 3.2.4) to the synthetic data. To that end, stochastic IP oversampling and parametric IP bagging, proposed by us (Section 3.2.2), require a distribution assumption for the main features $\tilde{\mathbf{X}}$. We always assume them to be normally distributed, even if the features in fact follow a Student's t, Poisson or Bernoulli distribution. We aim to find out how the algorithms get affected when assumptions are not met.

In fact the four different distribution scenarios meet the Gaussian assumption in decreasing order: The normal distribution trivially fulfills it. The t distribution is still continuous and symmetric so that the violation of the normality assumption may not get too severe. The Poisson distribution is discrete but approximately normal for $\lambda \geq 30$; however, in order to guarantee the normality assumption to be violated, we let $\lambda_i \in \{1, 2, 3, 4, 5\}$. The Bernoulli distribution cannot be seen as continuous and violates the normality assumption the most.

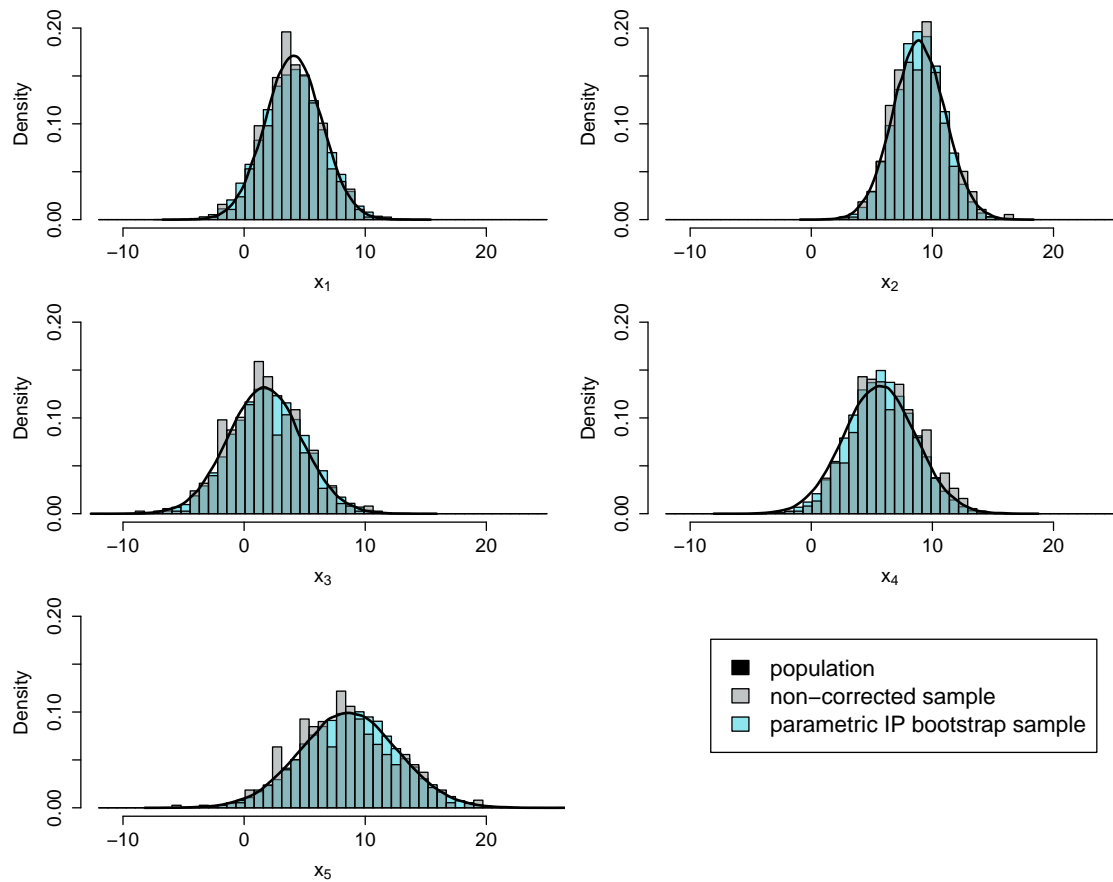


Figure 3.3: Distribution of variables from population, a biased sample and a generated sample corrected by parametric inverse-probability bootstrap for the normal distribution scenario. The corrected sample shows more similarity than the biased sample for the single variables (confirmed by Wilcoxon-Mann-Whitney-test).

Evaluation

We measure the performance of the different classifiers combined with the various correction approaches by the AUC. The AUC is appropriate especially in the context of sample selection bias since it does not require binary prediction (i. e. discretizing continuous risks by choosing a cut-off) and is unaffected by linear transformations of the predictions as only ranks are considered. Thus differences in performance should not be influenced by good or bad calibration of the prediction.

The goal of the comparison is to see whether correction approaches perform significantly better than not correcting. For each classifier, we fit a linear regression model with the

AUC as target variable and the correction approach as covariate. The latter variable is dummy-coded with 'no correction' as reference category. An approach is determined to differ significantly from the non-correction approach if its coefficient's t-test confidence interval does not contain zero. For all comparisons we use a level of significance of $\alpha = 5\%$.

Software

We used the statistical software R for all analyses [167]. More specifically, for building logistic regression models we used the R package *stats* [167], for random forest the R package *ranger* [190], and for naive Bayes the R package *e1071* [119]. The modified implementation of the SMOTE algorithm is based on the R package *smotefamily* [161]. We validated our results via ROC-analysis, using the R packages *pROC* [143] and *ROCR* [160].

3.3.2 Results

The simulation study yielded the following results, see also Figures 3.4 to 3.7: As expected, for every distribution scenario (see previous subsection) and all classifiers the performance of learning on the entire population was significantly better than learning without correction on the smaller biased learning data set. Also, for all classifiers and in all distribution scenarios, there was at least one correction technique that outperformed the non-correction approach (with two exceptions: logistic regression with additional interaction terms and naive Bayes, both in case of normally distributed main features).

However, there were differences between classifiers concerning the success of correction approaches. We start by contrasting logistic regression and the random forest as this comparison is of our primary interest:

The overall result for logistic regression (Figure 3.4) is that all correction approaches perform significantly better than non-correction. Exceptions are costing and modified SMOTE in the normal distribution scenario which on average performs better than non-correcting, but not significantly. For t-distributed and Poisson distributed features the difference between the performance of non-correction and the other approaches is more prominent than for the normal distribution scenario. In the Bernoulli case, this difference is

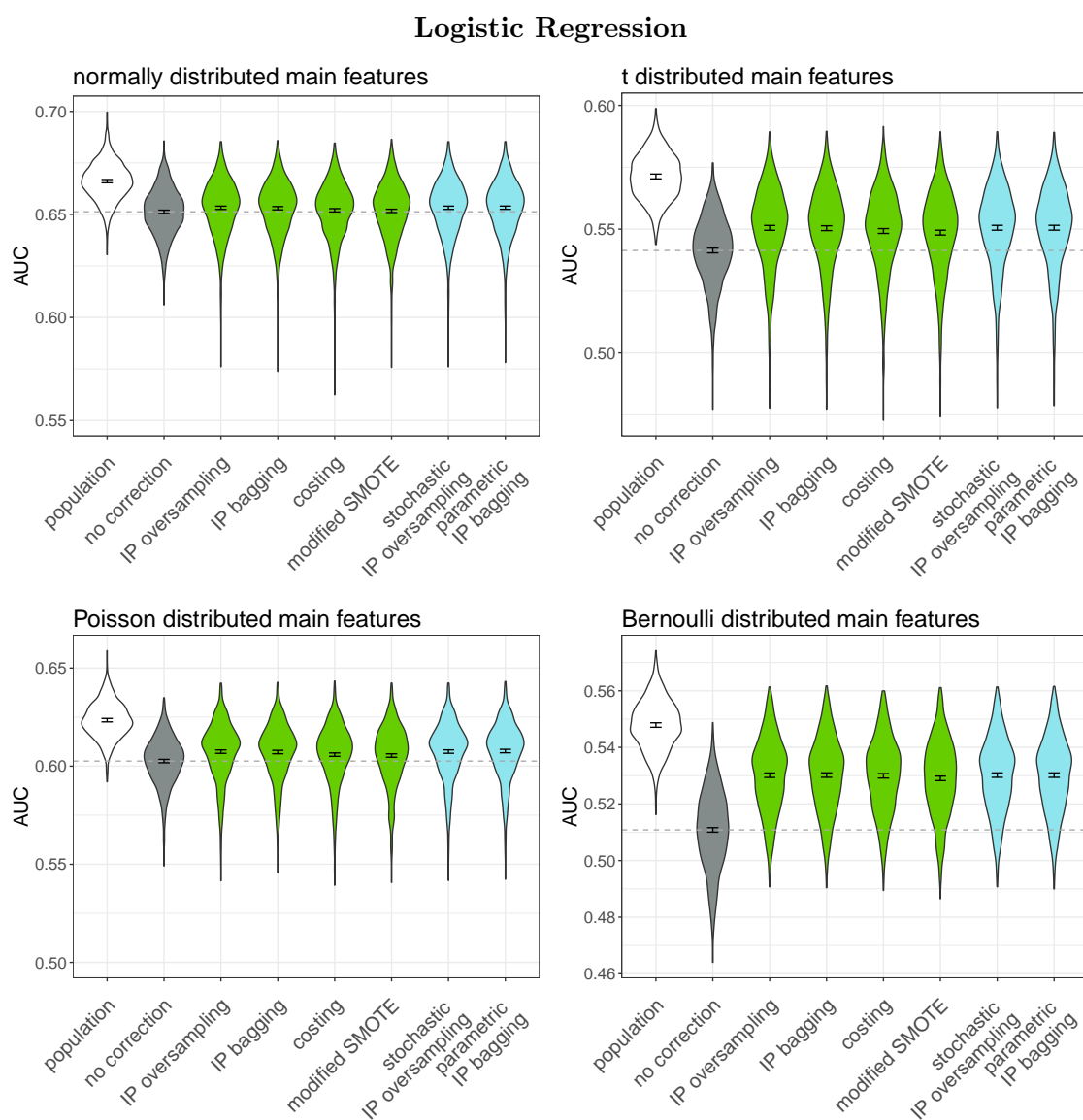


Figure 3.4: Performance of correction approaches in logistic regression, measured by AUC. We fit a linear model for the AUC as influenced by the correction method (dummy-coded, no correction as reference category). The graphic depicts 95% confidence intervals for the respective coefficients. The dotted line shows the intercept of the model, i.e. the mean AUC for no correction. The blue colored methods are newly proposed by us.

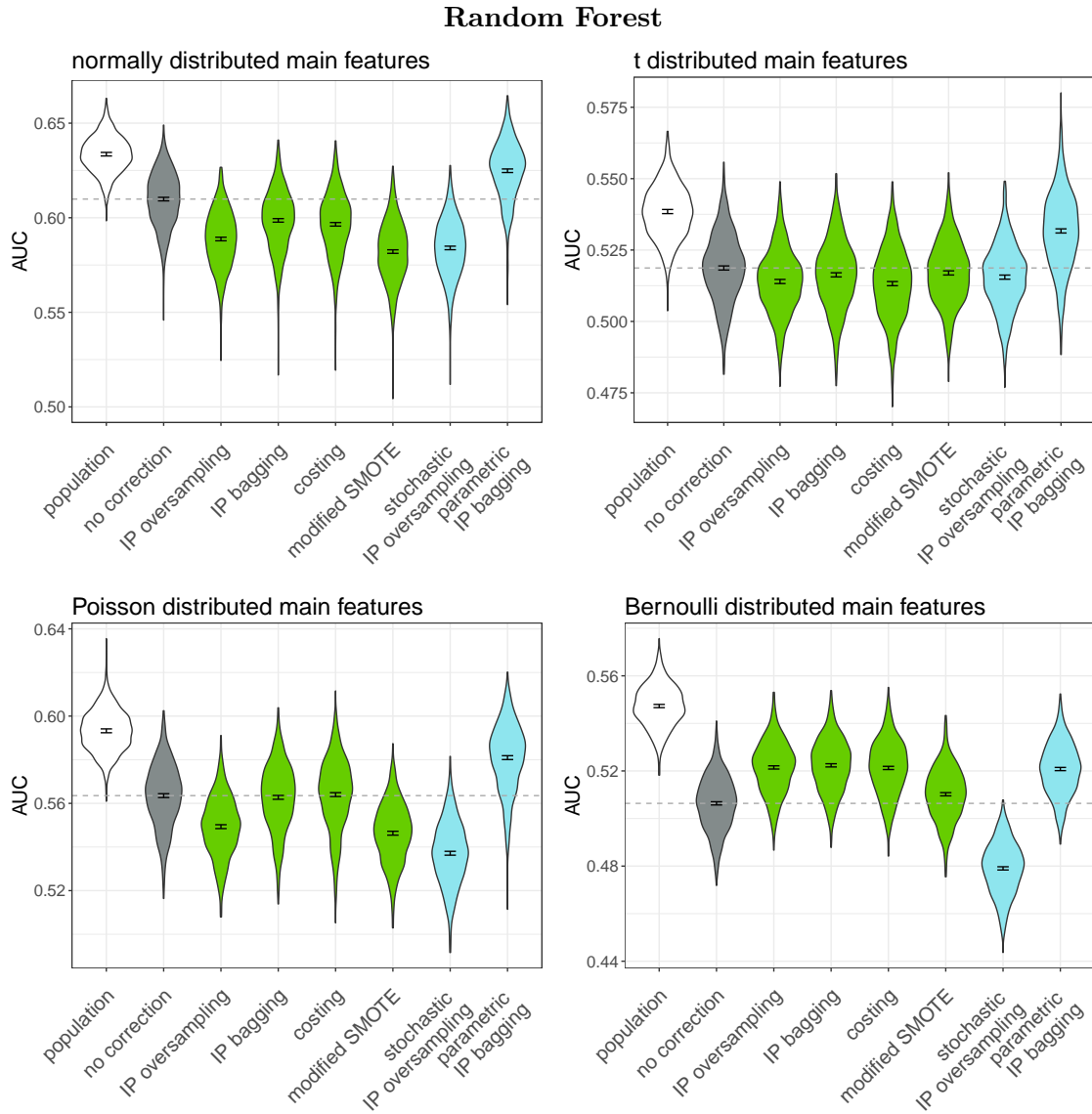


Figure 3.5: Performance of correction approaches in the random forest, measured by AUC. We fit a linear model for the AUC as influenced by the correction method (dummy-coded, no correction as reference category). The graphic depicts 95% confidence intervals for the respective coefficients. The dotted line shows the intercept of the model, i.e. the mean AUC for no correction. The blue colored methods are newly proposed by us.

the highest. Within each distribution scenario, the correction approaches perform similar to each other.

For the random forest the picture is rather different (Figure 3.5); only one correction approach performs significantly better than non-correcting: the parametric IP bagging proposed by us. In fact, for normally and t-distributed features all other correction methods perform even worse than non-correcting. In the Poisson scenario, they perform either worse than non-correction or equally fine (IP bagging and costing). Only in the scenario in which the assumption of having continuous main features (required by the approaches proposed by us) are not met at all, i. e. for the Bernoulli distribution, almost all correction approaches perform better than not correcting. An exception is stochastic IP oversampling proposed by us. This approach failed in all distribution scenarios for the random forest.

Correction approach	Properties according to Section 3.2.3			Sufficient performance	
	(i)	(ii)	(iii)	Logistic regression	Random forest
No correction	×	×	×	×	×
IP oversampling	✓	×	✓	✓	×
IP bagging	✓	✓	×	✓	×
Costing	✓	✓	×	(✓)	×
Modified SMOTE	✓	(✓)	✓	(✓)	×
Stochastic IP oversampling	✓	✓	✓	✓	×
Parametric IP bagging	✓	✓	✓	✓	✓

Table 3.1: Properties and performance of correction approaches for logistic regression and random forest. The properties are: (i) a correction attempt is made at all; (ii) the covariance structure of the learning data is attempted to be unbiased; (iii) learning is based on a data set containing a larger number n' of observations than the original stratified data set (see Equation 3.3). Criteria are fulfilled (“✓”), not clearly fulfilled (“(✓)”) or not fulfilled (“×”).

Table 3.1 summarizes the properties of the correction approaches (Section 3.2.3) together with the just described results. We label the performance of an approach to be sufficient if it results in a significant increase of the AUC as compared to the non-correction approach for the normal distribution scenario. Costing and modified SMOTE do not yield unambiguous improvements for logistic regression since their confidence intervals slightly overlap with the value under the null hypothesis. However, as we will see in Section 3.4, both approaches perform significantly better than non-correction on real data.

In order to obtain a more comprehensive picture of the benefit of correcting for sample

selection bias, we applied the correction methods in combination with two more classifiers, logistic regression with additional two-way interaction terms in addition to the linear terms and naive Bayes, leading to the following results:

Logistic regression with interaction terms yields a similar picture as standard logistic regression (Figure 3.6): All correction approaches perform similarly to each other. In the t- and Bernoulli scenario, again all correction approaches outperform the non-correction approach, except for costing for t-distributed features, which performs similar to non-correcting. For both the normal and the Poisson distribution, all correction approaches perform significantly worse than not correcting. An exception is parametric IP bagging: Similarly to the random forest case, only this method performs significantly better than no correction for the Poisson distribution scenario. For the normal distribution, the approach is the only one which does not perform significantly worse than the non-correcting approach.

For naive Bayes (Figure 3.7), again all correction approaches behave similarly as in logistic regression. Depending on the data distribution, correction approaches perform worse or better than non-correction. Especially in the normal distribution scenario the correction approaches are not successful.

3.4 Real data application

This section investigates the performance of the correction methods in a real data example. Other than in the synthetic data situation in the previous section, we do not know the true distribution of the entire population here. In order to still be able to evaluate the predictions appropriately, we chose a very large real data set from which we could extract a small stratified learning set and a large unbiased test set as described in the following.

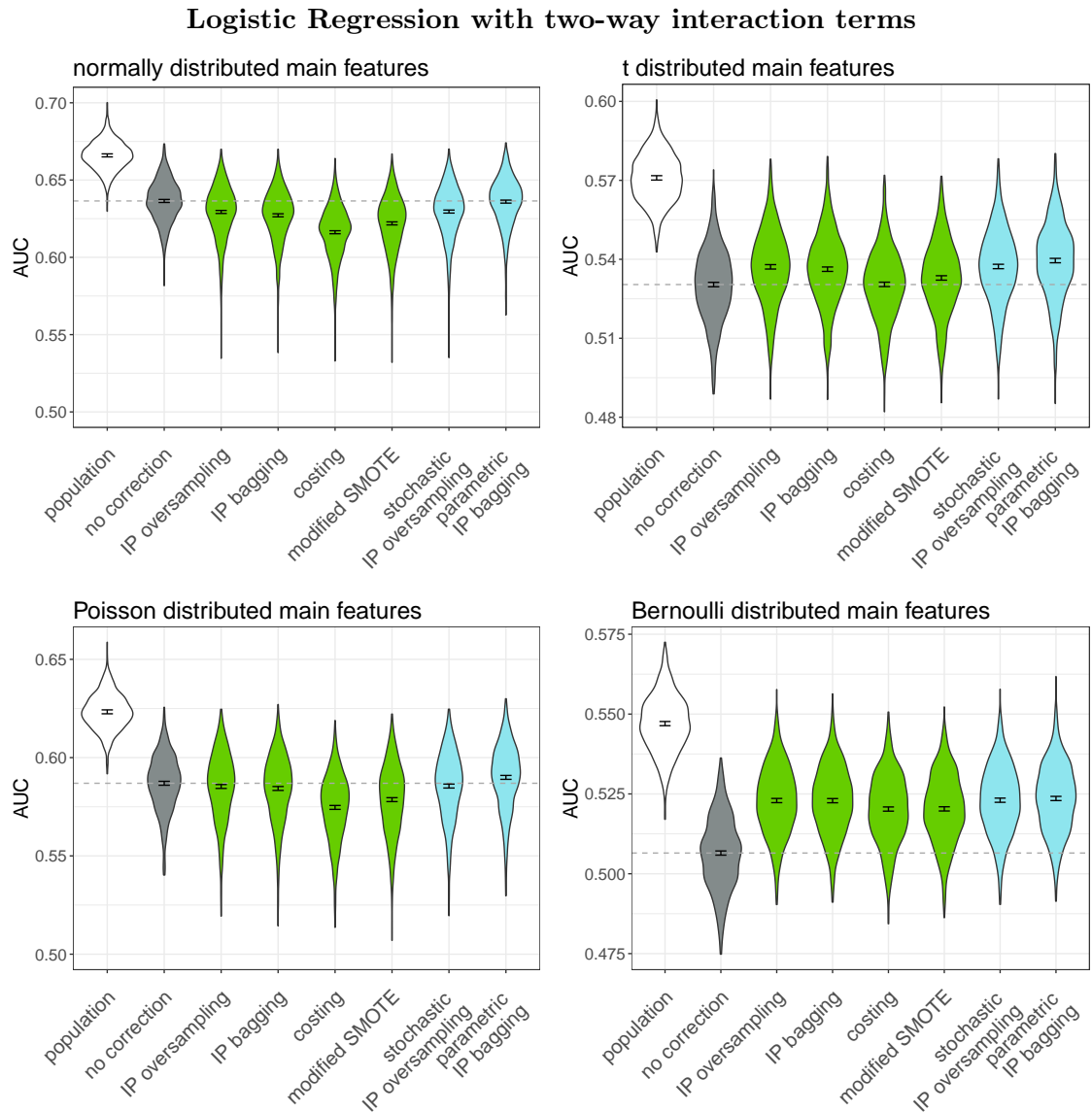


Figure 3.6: Performance of correction approaches in logistic regression with additional two-way interaction terms, measured by AUC. We fit a linear model for the AUC as influenced by the correction method (dummy-coded, no correction as reference category). The graphic depicts 95% confidence intervals for the respective coefficients. The dotted line shows the intercept of the model, i.e. the mean AUC for no correction. The blue colored methods are newly proposed by us.

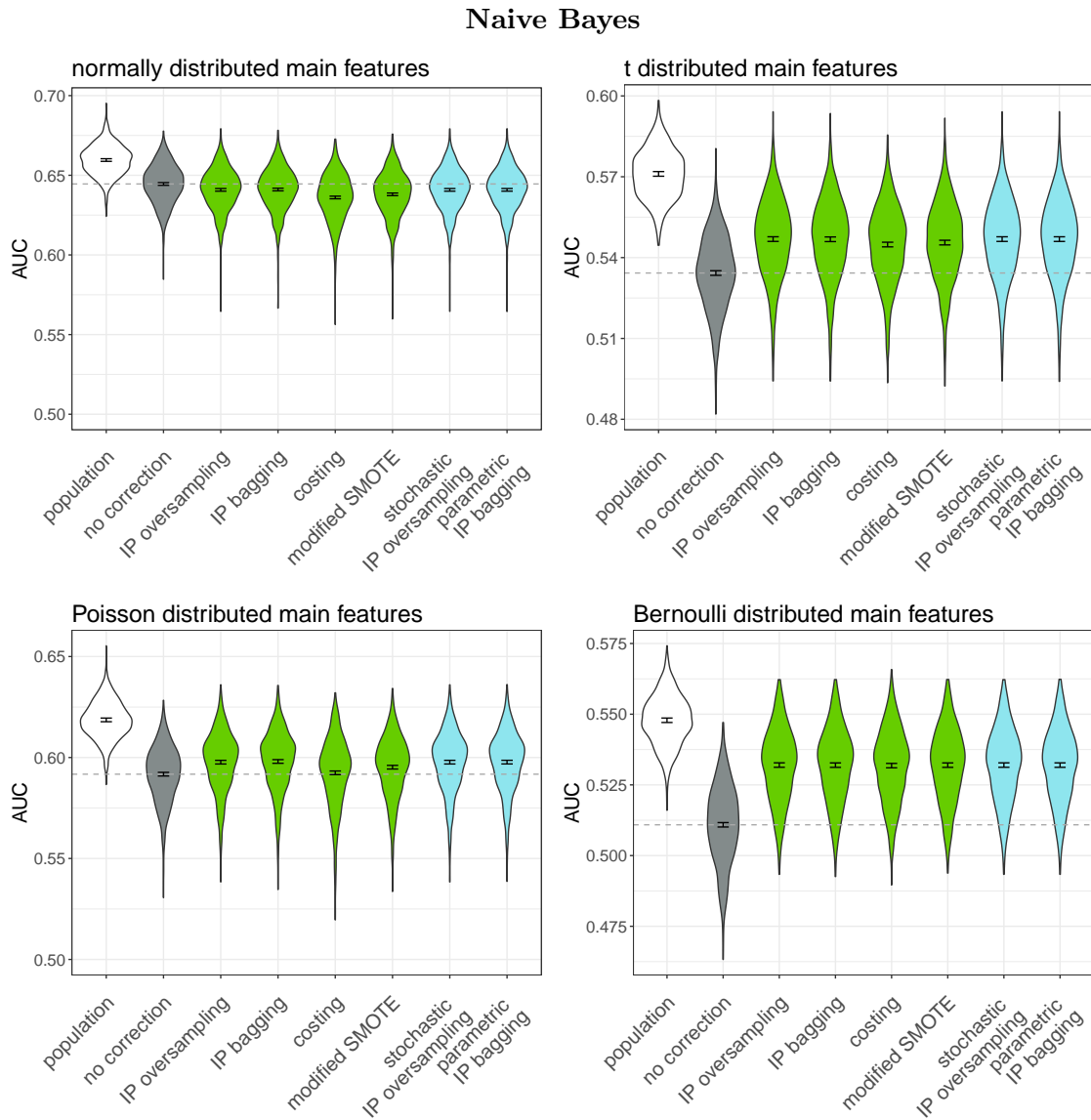


Figure 3.7: Performance of correction approaches in the naive Bayes classifier, measured by AUC. We fit a linear model for the AUC as influenced by the correction method (dummy-coded, no correction as reference category). The graphic depicts 95% confidence intervals for the respective coefficients. The dotted line shows the intercept of the model, i.e. the mean AUC for no correction. The blue colored methods are newly proposed by us.

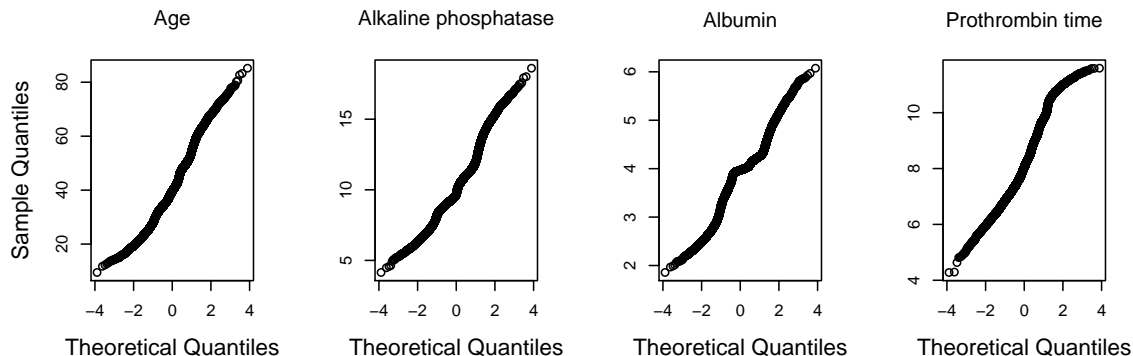


Figure 3.8: Normal quantile-quantile plots for main features $\tilde{\mathbf{X}}$ in real data set. For visualization purposes, we only displayed a random sample of 10,000 observations instead of the full data set of size 10^6 .

3.4.1 Design

Data

We evaluate the various prediction methods on the example of the *hepatitis* data set (data ID: 269, exact name: “BNG(hepatitis)”, version: 1) from OpenML [174]. It contains 10^6 observations of a binary outcome Y and 20 features. Y captures whether a hepatitis patient stayed alive and hence takes the categories *live* and *die*. We chose the binary variable *sex* as stratum feature X_e . From the remaining variables, we took into account the four continuous features *albumin*, *alkaline phosphatase*, *prothrombin time* and *age*, denoted by $\tilde{\mathbf{X}}$. These features were approximately normally distributed (partly after transformation, see the quantile-quantile plots in Figure 3.8) and strongly associated to the outcome.

Stratification process

We aimed to evaluate the prediction methods on data sets which underwent sample selection bias. We hence constructed a learning data set by performing a two-phase stratified random selection process on the *hepatitis* data set. To that end, we selected $n = 2000$ out of the 10^6 observations, enriching the outcome Y and the feature variable *sex*, denoted by X_e . Figure 3.9 shows the sizes of the four strata in analogy to Figure 3.1b. As test data set, we chose a subset of 10,000 observations from the hepatitis data set, disjoint to

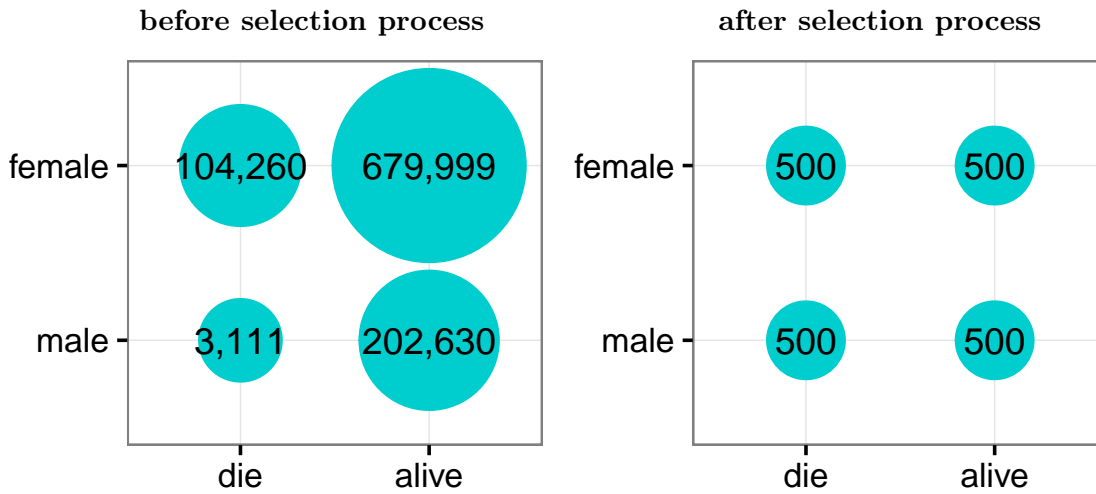


Figure 3.9: Cross table for the *hepatitis* data set before (left) and after (right) the selection process of a two-phase case-control study.

the learning data. We defined the first 10^6 observations (without the test data) as the population which served as reference learning data set as in the previous section.

3.4.2 Results

We trained all methods on the biased learning data and evaluated them on the unbiased test data. The resulting AUCs are compared by seven pairwise hypothesis tests according to [39]. We corrected for multiple testing via Bonferroni correction, i. e. set the threshold for p-values to $\alpha^* = 0.05/7 = 0.0071$.

The real data results confirm the findings from the simulation study. For logistic regression, all weighting approaches perform very similar, that was significantly better than the non-weighting approach and even comparable to learning on a large population (Figure 3.10a).

For random forest we obtain similar results as in the simulation study (Figure 3.10b): Only parametric IP bagging performs significantly better than the non-weighting approach. Costing and IP bagging perform insignificantly better, IP oversampling, modified SMOTE and stochastic IP oversampling perform significantly worse.

Also for logistic regression with interaction terms and naive Bayes we obtain results matching with the simulation study: the assumptions for normality are met only roughly for the real data, in which case the correction approaches all perform similarly and better than no correction (Figure 3.10d).

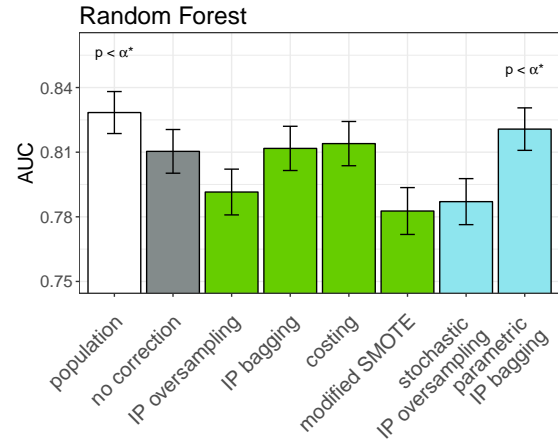
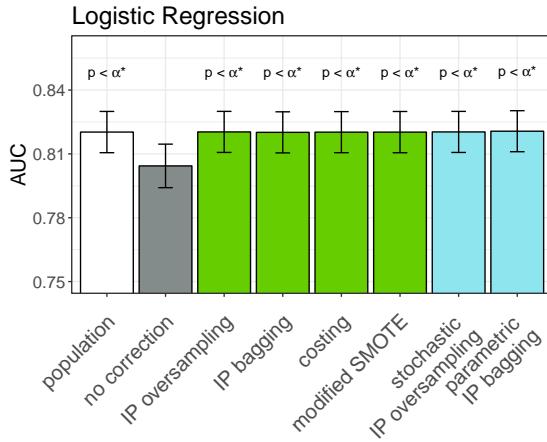


Figure 3.10a: Performance of logistic regression on real data. The graphic depicts 95% confidence intervals for the respective AUC value calculated and on the basis of [39]. All correction approaches perform similarly and significantly better than no correction (test by [39], $\alpha^* = 0.0071$).

Figure 3.10b: Performance of random forest on real data. The graphic depicts 95% confidence intervals for the respective AUC value calculated and on the basis of [39]. Only one correction approach, our novel parametric IP bagging, performs significantly better than no correction (test by [39], $\alpha^* = 0.0071$).

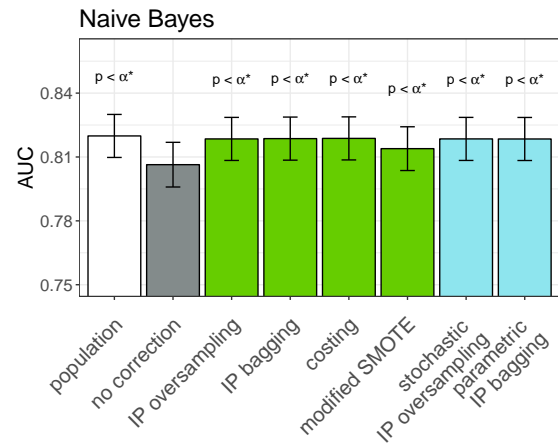
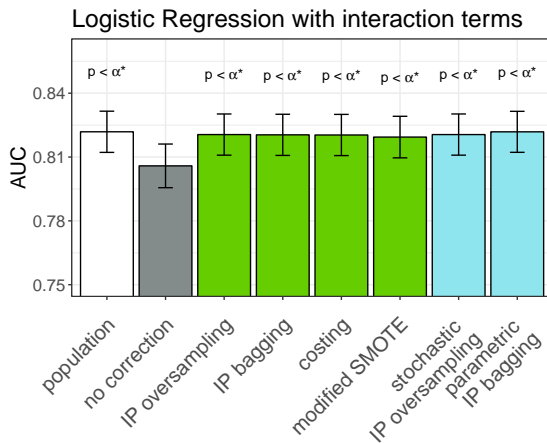


Figure 3.10c: Performance of logistic regression with all two-way interaction terms on real data. The graphic depicts 95% confidence intervals for the respective AUC value calculated and on the basis of [39]. All correction approaches perform significantly better than no correction (test by [39], $\alpha^* = 0.0071$).

Figure 3.10d: Performance of naive Bayes on real data. The graphic depicts 95% confidence intervals for the respective AUC value calculated and on the basis of [39]. All correction approaches perform significantly better than no correction (test by [39], $\alpha^* = 0.0071$).

3.5 Discussion and Conclusion

We investigated how to learn classifiers on stratified random samples as resulting from two-phase case-control studies. Here, our emphasis was on random forest classification since previous bias correction methods did not pay special attention to resampling-based classifiers. However, we studied a broad range of classification techniques. This work hence guides the choice of such approaches also for other classifiers. The methods are immediately applicable due to the implementations provided in our R package *sambias*.

State-of-the-art not always satisfactory — random forest requires novel method

Both our simulation study and the real data application show that prediction from biased on unbiased data sets can be improved if the stratification process is taken into account and corrected for. However, state-of-the-art correction approaches from classical statistics (IP oversampling, IP bagging, costing and modified SMOTE) do not yield the desired improvement for random forests. In fact, they can even lead to worse AUC values than those obtained when not performing any correction. From our two proposed approaches (stochastic IP oversampling and parametric IP bagging), on the other hand, the latter could always outperform the non-correction approach.

We were also interested in all correction approaches' success when employed in the context of logistic regression. It turned out that any method improves prediction on an independent data set as compared to no correction, and all correction techniques perform similarly.

Table 3.1 helps to explain the different behaviors of the two classifiers: Correction approaches are based on one or several of the principles (i) IP weighting, (ii) rebuilding the original covariance structure and (iii) increasing the number of learning observations as compared to the stratified sample. Obviously, weighting (Property i) should be applied in order to obtain any improvement in performance. Moreover, the covariance structure should be corrected for (Property ii) when applying a random forest. IP oversampling and partly modified SMOTE failed to fulfill this criterion. For logistic regression, in contrast, the covariance structure does not matter since point estimates of regression coefficients are not affected when the variance in the data is underestimated. Last, sample sizes

(Property iii) seem to matter more for random forests than for logistic regression. This is reasonable since too small sample sizes can restrict the range of the values of a feature and thus underestimate their variance leading to the same issue as for Property ii. This made IP bagging and costing perform poorly for the random forest. This leaves us with stochastic IP oversampling and parametric IP bagging, both proposed by us. However, although stochastic IP oversampling was designed to fulfill Properties i, ii and iii, we could not yield successful results for random forests.

The results for logistic regression suggest that all correction methods, except modified SMOTE, may approximately result in the identical classifier since their performances were very similar. This becomes clear to be the case, in fact, when the concepts of correction methods are compared: except for IP oversampling the correction approaches IP bagging, costing, stochastic IP oversampling and parametric IP bagging, are all based on aggregating resampled or bootstrapped data, so correspond to bagging methods. Bagging is an approach which can transform instable classifiers to stable classifiers. However, investigations have shown “when the classifier is rather stable, bagging is useless” [163]. The coefficient estimate for logistic regression is already the best linear unbiased estimator (BLUE) if the assumptions of the linear model are met, thus the estimator has minimum variance and is unbiased. Thus, bagging will in general not improve such a model.

In terms of correction approaches this holds as well: IP oversampling is identical to weighting observations in maximum-likelihood estimation for logistic regression. This implicates that all bagging correction approaches do not improve the model over the IP oversampling approach and thus not the prediction of that, but they work and are unbiased in terms of sample selection bias correction since weighting is incorporated.

Having compared correction methods in random forests and in logistic regression, one may conclude that the choice of parametric IP bagging is advisable whenever the distribution assumptions for this approach are met. In order to once more revise this conclusion, we investigated the behaviors of all correction approaches in two more classifiers, a logistic regression model with additional interaction terms and the naive Bayes classifier. For the logistic regression model with interaction terms, once again only the parametric IP bagging consistently outperformed the non-correction approach. For naive Bayes, all approaches performed similarly among each other, confirming the above stated rule.

Against our expectations, naive Bayes failed in the simulation study for the normal distribution scenario but did well for all other distributions. A generally unexpected result was the poor accomplishment of stochastic IP oversampling. It performed worse than

non-correction in several scenarios and was successful only in those situations where all other correction approaches were successful as well.

Parametric IP bagging — limitations and solutions

For a random forest, parametric IP bagging is an effective technique for prediction on an unbiased data set and can also be preferred for other classifiers. However, in this chapter we restricted our simulations and real data example to the case where the main features could be assumed to be roughly normally distributed (after transformation, if necessary) so that the assumption of a multivariate normal distribution was appropriate. The success of parametric IP bagging generally depends on meeting the assumptions about the distributions of the features. Hence, the method should be chosen with care. On the other hand, our simulations show that even in scenarios where assumptions are barely met (e. g. for Poisson distributed features), the approach still works. Clearly, one could also adjust the distribution family for the parametric bootstrap in parametric IP bagging. Even mixture distributions are conceivable e.g. for bimodal feature distributions.

So far, parametric IP bagging has not been designed for binary or categorical main features or combinations of different types. This could be done by subgrouping the corresponding categories (or combining categories in the case of several categorical features) and estimating parameters in each of the subgroups for the assumed distribution family analogously to what we did for the different strata. Again, one would draw parametric bootstrap samples within all subgroups and construct a new unbiased sample within the scope of parametric IP bagging.

Novel methodology recommended for stratified random samples

Even though our new approaches were developed for the random forest, they are generally tailored towards learning by any classifier and can be incorporated in other machine learning algorithms. Parametric IP bagging has shown to perform well even if theoretical assumptions are not met. It can be applied on any stratified random sample and is not restricted to two-phase case-control studies. More generally, it is suited for any sample suffering from sample selection bias where the stratum-features are categorical and the remaining features roughly follow a multivariate distribution from which parametric bootstrap samples can be drawn. For general classifiers, its performance is mostly com-

parable to that of other correction methods. Parametric IP bagging is the first correction method designed for the random forest and in that context clearly outperforms all other approaches.

With this project on sample selection bias in stratified random samples we have addressed one issue that often arises in studies caused by their special design where maximal power is desired. In particular we treated the case of training classifiers in two-phase case-control studies. The next chapter is an application on real data resulting from a study on childhood asthma where the study design exactly corresponds to a two-phase case-control study. There the enrichment of the exposure variable “farm” and the outcome variable “childhood asthma” cause the sample selection bias. Thus correction solutions will have to be taken into account when classifiers are trained and, in this case, also when they are validated, as also validation data is affected by sample selection bias.

3.6 Additional Material

Additional figures, code and data are available at <https://www.helmholtz-muenchen.de/index.php?id=47085> and in the appendix of this thesis.

Software in terms of the R-package *sambia* implements all correction approaches used in this thesis and is available on CRAN [103].

Chapter 4

Encountering big data: predicting childhood asthma risk by genetic and environmental variables

About two decades ago, the human genome project elicited much hope and fear about genome-wide testing. It was hypothesized that the resulting knowledge “may foretell future disease and [...] could be used to discriminate against or stigmatize a person” [37]. In the meantime, international consortia bringing together more than 10,000 cases [115, 123] have discovered various genetic determinants of childhood onset asthma, a paradigm of a polygenic disease with additional environmental determinants [45], which may even interact with the underlying genotype [131]. With these genetic and environmental determinants, much of the variance in population studies can be explained. But what does this imply for an individual? Do these determinants “foretell future disease” [37] in individuals?

In their meta-analysis of genome-wide association studies (GWAS) on childhood onset asthma, Moffatt and colleagues actually reported low predictive values with a very modest AUC of 0.58, with 0.5 marking no predictive value at all [123].

This modest prediction quality conflicted with the strong hereditary background postulated from twin studies [43]. On the other hand, the relatively weak association of asthma with family history in the GABRIELA study [80] contrasts with the twin studies and points towards interactions with environmental exposures likewise shared by twins.

In search of gene-environment interactions for childhood asthma exposures related to farming have been explored with inconclusive results. This is illustrated by the interaction of farm milk consumption with the innate immunity receptor CD14 [17], which later was invalidated in a well-powered genome-wide interaction study [49].

The impact of genetic loci on clinical phenotypes is typically tested or predicted univariately. However, with about 0.6 to 0.7 million independent loci in the human genome [42], the multiplicity of univariate tests severely reduces the overall statistical power. This is a conceptual limitation of the classical test theory and can hardly be overcome by merely increasing case numbers. In risk prediction modeling, likewise, univariate models limit the predictive power as they leave out potentially important dependencies between loci. The concurrence of all loci is exploited most effectively by incorporating them into one multivariable model. However, the large amount of variables poses a computational challenge and often discourages researchers from multivariable analyses. The aim of the present study was to explore novel statistical tools, which consider predictor variables integratively rather than separately. We hypothesized that these methods would substantially improve individual-level disease prediction based on genetic information and interactions with environmental variables.

This chapter is in parts identical with the following manuscript:

[101]: Norbert Krautenbacher, Michael Kabesch, Elisabeth Horak, Charlotte Braun-Fahrlander, Jon Genuit, Andrzej Boznanski, Erika von Mutius, Fabian J Theis, Christiane Fuchs, and Markus J Ege. Predicting childhood asthma risk by genetic and environmental variables. *submitted*, 2018

4.1 Data collection

The following data collection methods have been performed by the clinical partners involved in this chapter's project or members who were involved in the study conduct.

4.1.1 Population and questionnaires

The participants of this analysis were enrolled in the Austrian, Swiss, and German arms of the cross-sectional GABRIEL Advanced Studies (GABRIELA) [65]. By a stratified random selection process (Figure 4.1), informative children were enriched in order to gain maximum power for association analyses in genome-wide data [49]. The selection process

corresponded to a two-phase case-control study similar to the situation investigated in Chapter 3. These 1,707 GABRIELA participants were also included in previous meta-analyses of asthma [123].

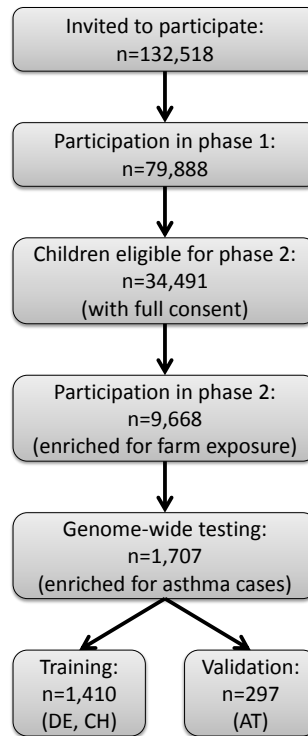


Figure 4.1: Participant flow in the GABRIELA study

The questionnaires contained items on individual and family health, socioeconomic background, and farm-related exposures. According to the German ISAAC definition [182], childhood asthma was defined in school children as a physician diagnosis of asthma at least once or of asthmatic bronchitis at least twice. If a child lived on a farm run by the family, the child was termed “farm child” ($n = 483$), and “non-farm child” ($n = 1,224$) otherwise. Other farm-related exposures were related to raw milk consumption or contact with animals or animal feed. Those variables were included either as exposure in the first years of life or as exposure during the past 12 months.

For external validation, the models were trained on the Swiss and the two German arms

Table 4.1: SNPs reported for childhood asthma in the GWAS catalog [114]

SNP	Region	Reported genes*	Comments
rs4658627	1q44	C1orf100	chromosome 1 open reading frame 100
rs9815663	3p26.2	IL5RA	interleukin 5 receptor subunit alpha
rs2705520	3q13.2	ATG3	Autophagy Related 3
rs17033506	3p22.3	(intergenic)	ARPP21: cAMP Regulated Phosphoprotein 21
rs9823506	3q12.2	ABI3BP	ABI Family Member 3 Binding Protein
rs6871536	5q31.1	RAD50	RAD50 Double Strand Break Repair Protein
rs1295686	5q31.1	IL13	Interleukin 13
rs2473967	6q21	(intergenic)	LOC105377956; LOC105377953
rs6967330	7q22.3	CDHR3	Cadherin Related Family Member 3
rs9297216	8p12	(intergenic)	LOC105379365
rs16929097	9p23	(intergenic)	TYRP1 (Tyrosinase Related Protein 1)
rs11141597	9q21.33	(intergenic)	LOC105376124 / GAS1
rs928413	9p24.1	IL33	Interleukin 33
rs7927044	11q24.2	(intergenic)	LOC107984373; LOC387820
rs7328278	13q13.3	(not reported)	DCLK1 (Doublecortin Like Kinase 1)
rs10521233	17p12	(intergenic)	LOC105371544; LOC107985014
rs2305480	17q21.1	GSDMB	Gasdermin B
rs3894194	17q21.1	GSDMA	Gasdermin A
rs7216389	17q21.1	ORMDL3	ORMDL sphingolipid biosynthesis regulator 3

* Genes are reported by authors of original publications [114]. If no genes are reported, mapped genes are given in the comments column.

of GABRIELA ($n = 1,410$) and validated in the independent population of the Austrian GABRIELA arm ($n = 297$); the relation of the sizes of training and validation datasets was chosen for an optimal trade-off of variance and bias [74]. In addition, $n = 928$ children of the prospective PASTURE birth cohort served for external validation of the final model in a cohort design.

4.1.2 Genotyping

Genotyping was performed with the Illumina Human610 quad array (Illumina Inc, San Diego, Calif, <http://www.illumina.com>), and quality was assessed as described previously [123]. SNPs were identified by linkage disequilibrium patterns from the HapMap CEU SNP panel version 2 and the 1000 Genomes pilot 1 release [64]. Candidate SNPs (Table 4.1) were defined as SNPs included in the GWAS catalog for childhood onset asthma [114].

4.2 Computational and statistical analysis

All statistical analyses were performed with R software [167]. We employed genotyped SNPs and imputed SNPs, obtained from HapMap CEU SNP panel version 2 with the use of Markov chain-based haplotyper [110]. SNPs were filtered for imputation quality: the so called *estimated r^2* or *Rsq* is the estimated correlation between imputed and true genotypes. We removed SNPs with $Rsq < 0.30$. In addition, SNPs with a low minor allele frequency ($MAF < 0.05$), i.e. SNPs where the less frequent allele occurred too rarely, were removed. Using the R package SNPRelate, we pruned for linkage disequilibrium by removing SNPs within a $5 \cdot 10^5$ SNP window that had $r^2 > 0.95$ [191]. By this we reduced the imputed SNP data from 2,543,887 SNPs to 744,908 SNPs.

Imputing missing values

Environmental variables (Supplemental Table A.1) had less than 25% missing values; missing values of existing variables were imputed by multiple imputation with the MICE algorithm using the R package MICE which we described in Chapter 2 Section 2.2.3. Continuous variables were imputed by predictive mean matching, binary variables by logistic regression, multicategorical variables by a multinomial logit model. We used five imputations and kept the required multiple imputation steps through all our analyses according to Rubin [148] (see also Section 2.2.3).

Response

Apart from the main response variable doctor-diagnosed asthma, we investigated the asthma phenotype in a more unambiguous way. To that end, we left out subjects who reported current wheeze or used asthma sprays but were not diagnosed with asthma (sensitivity analysis). By this definition, 166 of the 1707 subjects dropped out.

In a further investigation, we defined children to have asthma irrespectively of bronchitis. Children without doctor-diagnosed asthma were treated as controls. By this definition 4 of the 1707 subjects dropped out due to missing data on asthma diagnosis.

Prediction models

We compared multivariable approaches to a state-of-the-art strategy for developing a prediction model from GWAS: There are several ways of selecting SNPs univariately and building prediction models from it. According to an extensive investigation of these ways [191], the best strategy is a one-phase procedure, which we chose using all training data to select SNPs and to estimate the coefficients for the SNPs. We ranked the p-values starting with the smallest and included the best 100 SNPs for building a prediction score as recommended in the literature [191], yielding the highest AUCs according to their investigations: Marginal multiple simple logistic regression models were fitted for each SNP j , i.e. $P(y_i = 1|x_{ij}) = (1 + \exp(-(\beta_0 + \beta_j x_{ij})))^{-1}$. We then built a score S_i for subject i from the sum of the univariately estimated coefficients of the 100 top SNPs:

$$S_i = \sum_{j=1}^{100} x_{ij} \hat{\beta}_j.$$

In addition to classical GWAS performing an association test univariately for each single SNP [123, 191], we incorporated all variables at once in multivariable statistical learning models with the following regularization methods: the least absolute shrinkage and selection operator (LASSO) and elastic net (cf. Section 2.3.3) as our goal was to select important variables from many candidate variables, so that we used those penalty terms that penalize non-influential coefficients to zero. Further, in order to incorporate all modalities appropriately and avoid issues that can arise for multi-omics data (cf. Chapter 1), we implemented the integrative L1-penalized regression with penalty factors (IPF-LASSO) (cf. Section 2.3.4). As a further state-of-the-art machine learning method we used the random forest (cf. Section 2.3.6).

Each of the mentioned multivariable classification models was applied to the determinants demographics, environment, family-history, and genetics, each separately and then step by step on several to all determinants at once including a further modality incorporating selected interaction terms between the modalities.

Sample selection bias in data at hand

Since the data were not taken by pure random sampling but by a stratified random selection process (details in Ege et al. [49]), sample selection bias occurred [100]. Statistical analysis had to take this into account. From the selection process, selection probabilities

were available so that appropriate information from an unbiased large population (34,4391 children obtained after phase 1) could be taken into account. This was incorporated in correction approaches in the learning and the validation procedure as elaborated in the following.

To correct the point estimates for the effects for sample selection bias in logistic regression, the observations were weighted by inverse-probability weights w_i [100] which were given after the stratified random selection process [162]. Then the modified log-likelihood was given by

$$\ell_{mod}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^n w_i (y_i (\beta_0 + \mathbf{x}'_i \boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \mathbf{x}'_i \boldsymbol{\beta}))).$$

Since this is a modification of maximum likelihood estimation in the common sense, this function is called pseudo-maximum-likelihood [162]. This correction was applied to all classifiers based on logistic regression. The approach results in the same coefficient estimates as IP oversampling (cf. [100]/Chapter 3).

As this correction still gives biased estimates for the standard errors of the regression coefficients, for the prediction model based on GWAS and ranking p-values the standard errors were adjusted: for p-value calculation, the standard errors of $\hat{\boldsymbol{\beta}}$ were adjusted by design-based standard errors with approximation via Taylor series with the R-package *survey* [113].

Random forests, however, as an ensemble of many classification trees can be affected by the so-called imbalanced data classification problem [35]. In general, state-of-the-art correction methods for sample selection bias would cause such imbalanced data in the present study; our investigations have shown that most correction methods indeed do decrease performance of the random forest if the data at hand is sampled from such a stratified random sample ([100]/Chapter 3). The novel approach parametric inverse-probability bagging which has shown to be successful for random forests in Krautenbacher et al. [100]/Chapter 3 was not suitable for the data at hand: especially SNP data, seen as continuous variables, typically have skew distributions rather than being eligible for being approximated by a multivariate normal distribution — the distribution should be either symmetric (and to be approximated by a multivariate normal distribution) or be a Bernoulli distribution, but none of it is the case. Thus, for this method we went without correcting for sample selection bias in the learning process; otherwise the classes of the response would get strongly imbalanced. However, for comparison we applied parametric IP bagging and show results in the supplement (Figure A.11).

Model selection and validation

Model selection and 5-fold cross-validation was performed on the 1,410 Swiss and German participants. The best models were then externally validated in the 297 Austrian participants and additionally in the PASTURE birth cohort. As a metric for model comparison, we applied the AUC with a bootstrapped 95%-confidence interval [54]. If not indicated otherwise, AUC values refer to random forest models.

Variable importance

In order to identify the most important variables for prediction, we considered those variables that were selected by the most successful prediction model. In most of the analyses the random forest turned out to perform best, however, for the best model on farm children also IPF-LASSO nearly performed as well as the random forest. In the latter analyses we investigated both models.

In case of random forests we used Altmann's method (cf. Chapter 2). For variable importance in genome-wide applications, however, we used a further version based on cross-validation instead of OOB (out-of-the-bag) observations, proposed by [89]. This version works well for high-dimensional data (but was too unstable for the low-dimensional analyses) but is a much faster implementation of a random forest variable importance measure. For the latter we took the multiple testing issue into account, see also the Supplemental Section A.2.1.

For the IPF-LASSO we determined the most important variables as follows. Any version of LASSO per definition only selects variables contributing to good prediction. Hence, we interpreted the selected variables as the important ones. We determined the degree of importance of these variables as the size of the highest penalty λ for which the corresponding coefficient still remained non-zero. This is plausible since the more predictive a variable, the stronger one can penalize it without its coefficient being set to zero by the LASSO-procedure.

4.3 Correcting and comparing losses for sample selection bias

In this section we will propose how to correct a loss, here in terms of AUC, for sample selection bias when classifiers have to be validated. Weights have been incorporated for calculating the loss [38] and been implemented for the AUC [57]; however, no unbiased confidence intervals have been provided.

Further, we will propose a test of significance for comparing two losses, again here in terms of AUC. Such a test has been proposed [39] for comparing AUCs; however, it will give biased results under sample selection bias. In our bootstrap-based [46] version inverse-probability weights are incorporated in order to correct for sample selection bias.

Correcting confidence intervals for AUC by bootstrap using selection probabilities

Sample selection bias does not only affect the training of a model but also its evaluation. In addition, confidence intervals for the AUC are desirable but would be biased without correcting for sample selection bias. Hence, both the evaluation and the confidence intervals should be adjusted, e.g. by weighting [38]. However, an established evaluation weighting approach given for certain loss functions [38] is not directly applicable to the AUC. Therefore, we propose a different approach resolving both issues: We perform weighting by the use of bootstrap [46]. For the predicted risk $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$ and the corresponding true response $y = (y_1, \dots, y_n)$, the corrected AUC is then given by

$$\frac{1}{B} \sum_{b=1}^B L_{AUC}(\hat{y}^b, y^b)$$

where the pair (\hat{y}^b, y^b) corresponds to the b -th bootstrap sample, $b \in \{1, \dots, B\}$, which is built by resampling n elements with replacement from (\hat{y}, y) , using selection probabilities proportional to w_i for observation i . L_{AUC} denotes the loss function, which here corresponds to the AUC. We construct a percentile-confidence interval

$$[L_{AUC,2.5\%}(\hat{y}^b, y^b), L_{AUC,97.5\%}(\hat{y}^b, y^b)]$$

with $L_{AUC,q}(\hat{y}^b, y^b)$ denoting the empirical q -quantile of the B bootstrap values $L_{AUC}(\hat{y}^b, y^b)_{b=1, \dots, B}$. For all our analyses in this chapter, we chose $B = 10,000$.

A bootstrap test for pairwise AUC comparison using selection probabilities

Similarly to the corrected confidence intervals introduced above, we implemented a test based on selection probabilities for the pairwise comparison of two AUC values when validated on the same data. Let $\hat{y}^{(1)} = (\hat{y}_1^{(1)}, \dots, \hat{y}_n^{(1)})$ be the predicted risk by a first classifier, $\hat{y}^{(2)} = (\hat{y}_1^{(2)}, \dots, \hat{y}_n^{(2)})$ the predicted risk by a second classifier and again $y = (y_1, \dots, y_n)$ the corresponding true response. We regard the corrected difference of AUCs

$$\frac{1}{B} \sum_{b=1}^B D_{AUC}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)$$

with

$$D_{AUC}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b) := L_{AUC}(\hat{y}^{(1),b}, y^b) - L_{AUC}(\hat{y}^{(2),b}, y^b)$$

where the pair $(\hat{y}^{(k),b}, y^b)$ for classifier $k \in \{1, 2\}$ corresponds to the b -th bootstrap sample, $b \in \{1, \dots, B\}$, again, taken by using selection probabilities proportional to w_i . We construct a percentile-confidence interval for this difference by

$$[D_{AUC,2.5\%}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b), D_{AUC,97.5\%}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)]$$

with $D_{AUC,q}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)$ denoting the empirical q -quantile of the B terms $D_{AUC}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)_{b=1, \dots, B}$, again obtained via bootstrap with selection probabilities proportional to w_i . As before, we chose $B = 10,000$. We consider two classifiers to perform significantly different — i.e. we reject the null hypothesis H_0 : “AUC for classifier 1 is equal to AUC for classifier 2” — if the confidence interval does not overlap with zero. Analogously, one can test H_0 : “AUC for classifier 1 is less than or equal to AUC for classifier 2”, if AUC for classifier 1 is expected to be at least as good as classifier 2. The corresponding one-sided percentile-confidence-interval is

$$[D_{AUC,5\%}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b), 1].$$

4.4 Results

The $n = 850$ cases and $n = 857$ controls included in the present analyses differed with respect to sex, family history of asthma and atopy, and various farm-related exposures (Table 4.2, Supplemental Table A.1).

Table 4.2: Potential determinants of asthma. A selection of variables included in demographics, family history or environment are shown. Sex, family history of asthma and atopy, consumption of farm milk, contact with cows or straw and living on a farm show significant association to childhood asthma.

Characteristic	Cases (%) n=850	Controls (%) n=857	p-value*
female sex	39.70%	49.40%	0.002
age [§]	8.32 (0.06)	8.19 (0.06)	0.15
body mass index [§]	17.11 (0.11)	16.99 (0.11)	0.375
family history of atopy	70.00%	49.40%	<0.001
family history of asthma	30.06%	12.40%	<0.001
living on a farm	9.00%	13.60%	<0.001
at least two siblings	0.42 (0.02)	0.45 (0.02)	0.374
high parental education	27.30%	28.80%	0.633
maternal smoking during pregnancy	12.40%	8.50%	0.037
consumption of farm milk during past 12 months	13.40%	19.40%	<0.001
consumption of farm milk in first year of life	6.20%	11.80%	<0.001
consumption of farm milk (pregnancy to age 3yrs)	20.70%	27.60%	<0.001
contact with cows (past 12 months)	12.90%	16.60%	0.02
contact with cows (pregnancy to age 3yrs)	14.60%	20.30%	<0.001
contact with straw (past 12 months)	15.70%	21.10%	0.009
contact with straw (pregnancy to age 3yrs)	12.40%	16.20%	0.009
contact with hay (past 12 months)	29.70%	33.50%	0.145

* p-values based on Fisher's exact test or, in case of continuous variables, Wilcoxon tests
[§] mean and standard error of mean

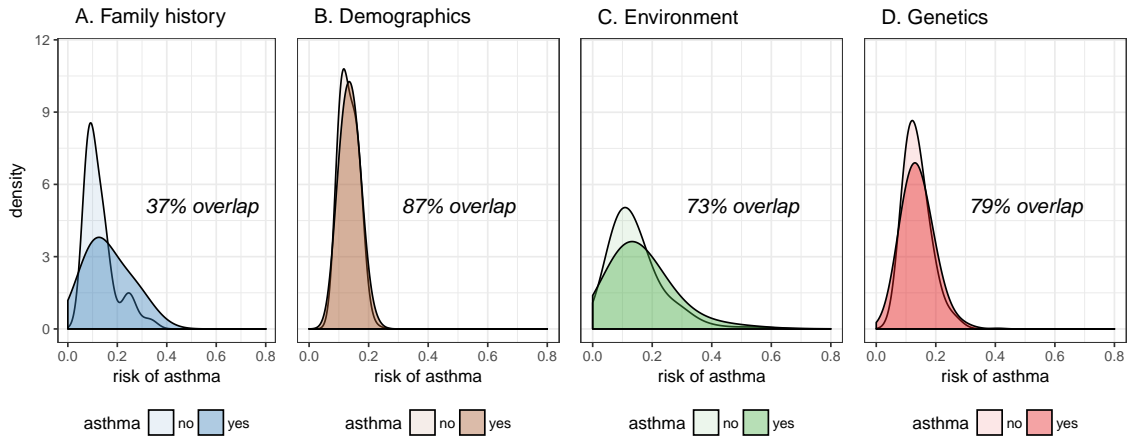


Figure 4.2: Risk of childhood asthma by environmental exposure, family history, and genetics. Individual risks are assessed by weighted logistic regression in a 5-fold cross-validation procedure based on (A) the four variables for family history; (B) the three demographics variables sex, age, and body mass index, (D) the environmental variables listed in Table A.1 in the supplement, and (D) the 19 candidate SNPs listed in Table 4.1. Empirical density functions are estimated by kernel density estimation. The distribution of the individual asthma risk differed clearly between cases and controls in all models as confirmed by Kolmogorov-Smirnov tests with all $p < 10^{-6}$. The overlap in density mass between cases and controls is given as percentage.

The distribution of the individual asthma risk estimated by kernel density estimation was assessed by separate weighted logistic regression models for family history of asthma, demographics, environmental variables, and genetics (Figure 4.2). Though the distribution of the individual asthma risk differed between cases and controls in all models, there was a substantial overlap of the asthma risk distribution as determined by the area under the density functions (density mass) between both groups, rendering prediction of asthma risk on an individual level very difficult. The most pronounced difference in asthma risk between cases and controls was found for family history with only 37% of density mass overlap.

Given the inadequate separation of asthma risk between cases and controls, we explored various multivariable learning approaches (multivariable logistic regression with LASSO penalty, multivariable logistic regression with elastic net penalty, and random forest) to improve the discriminatory power of the prediction. When assessing groups of variables, i.e. family history, demographics (sex, age and BMI), environment, and genetics separately, the various methods did not differ with respect to prediction quality (Figure 4.3, upper panel).

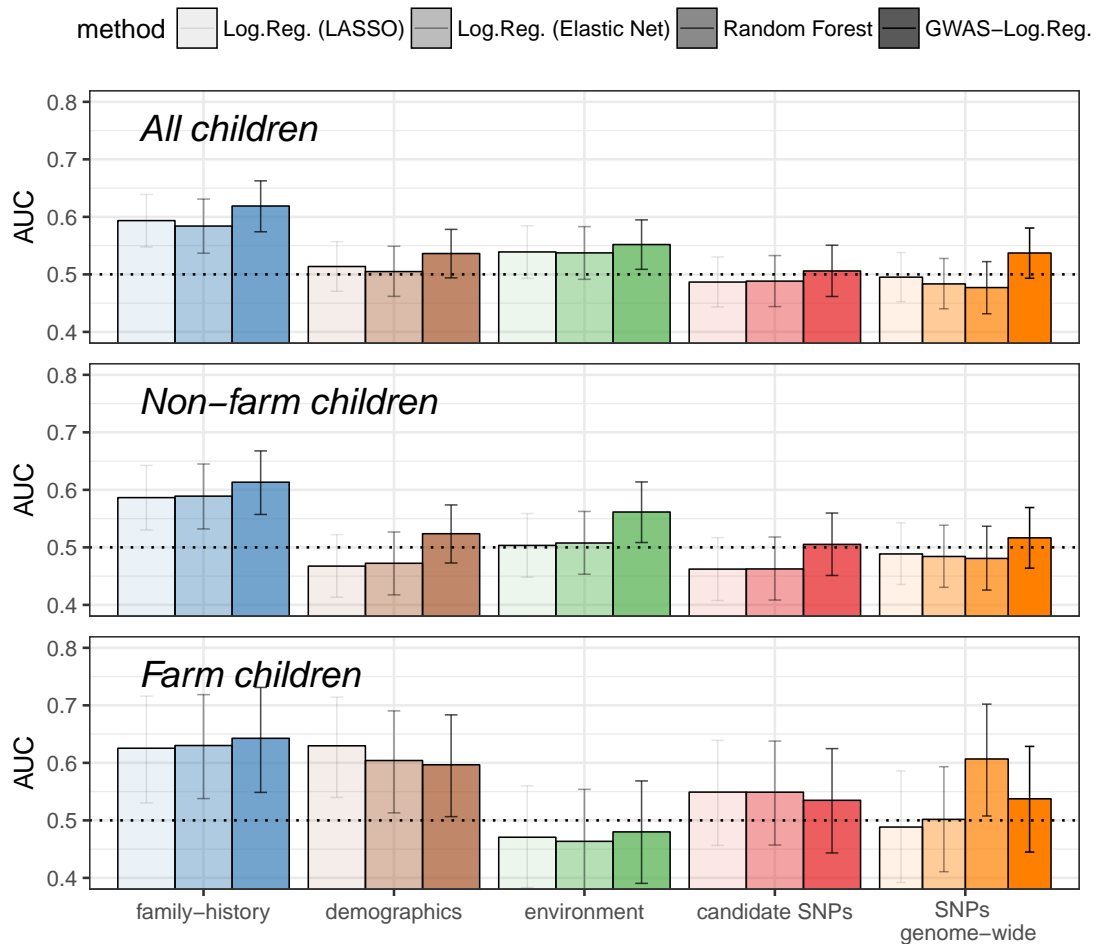


Figure 4.3: Comparison of prediction performance for different groups of predictors and statistical methods. Prediction performance of the variable groups family-history, demographics, environment, candidate SNPs (Table 4.1), and genome-wide SNPs on a stand-alone basis. As statistical methods, we used multivariable logistic regression with LASSO penalty, multivariable logistic regression with elastic net penalty, the random forest and multiple logistic regression models. The AUC is calculated as mean over 5 imputation data sets with 95% confidence intervals constructed by bootstrap using selection probabilities. The dotted line at 0.5 corresponds to the AUC-value where a prediction model classifies cases and controls not better than at random.

Again, family history was the best predictor of childhood asthma with an AUC value of 0.62 [0.57-0.66] in the random forest model. All other groups of variables did not predict better than by chance except for environmental variables in the random forest model (AUC=0.55 [0.51-0.59]). When stratifying prediction models for the two major study groups, i.e. farm children and non-farm children, we noted that the joint model was driven by the

non-farm children (Figure 4.3, middle panel), who accounted for about two thirds of the population. For farm children, however, different prediction models emerged: instead of environmental variables, demographics and genome-wide SNPs (AUC=0.61 [0.51-0.70]) predicted significantly (Figure 4.3, lower panel). Of the 744,924 GWAS SNPs about 3700 SNPs were estimated to be significant in the genome-wide prediction model (see Supplemental Figure A.7). Among asthma cases many more SNPs differed significantly between farm and non-farm children as compared to non-cases (see Supplemental Figure A.8).

In an effort to explore combined effects of all predictors, we sequentially complemented prediction by family history with demographics, with environmental variables or candidate SNPs or both, and finally by interaction terms for the SNPs and the other variables. For all groups of variables, random forest and IPF-LASSO performed much better than simple LASSO (Figure 4.4, upper panel) and the other techniques. Prediction by family history was significantly (Table 4.3) improved by demographics and environmental variables (AUC= 0.65 [0.61-0.70]) or, in case of farm children, by demographics and candidate SNPs (AUC= 0.70 [0.62-0.78]), whereas GWAS SNPs and interaction terms did not further improve prediction quality (Figure 4.4, lower panel).

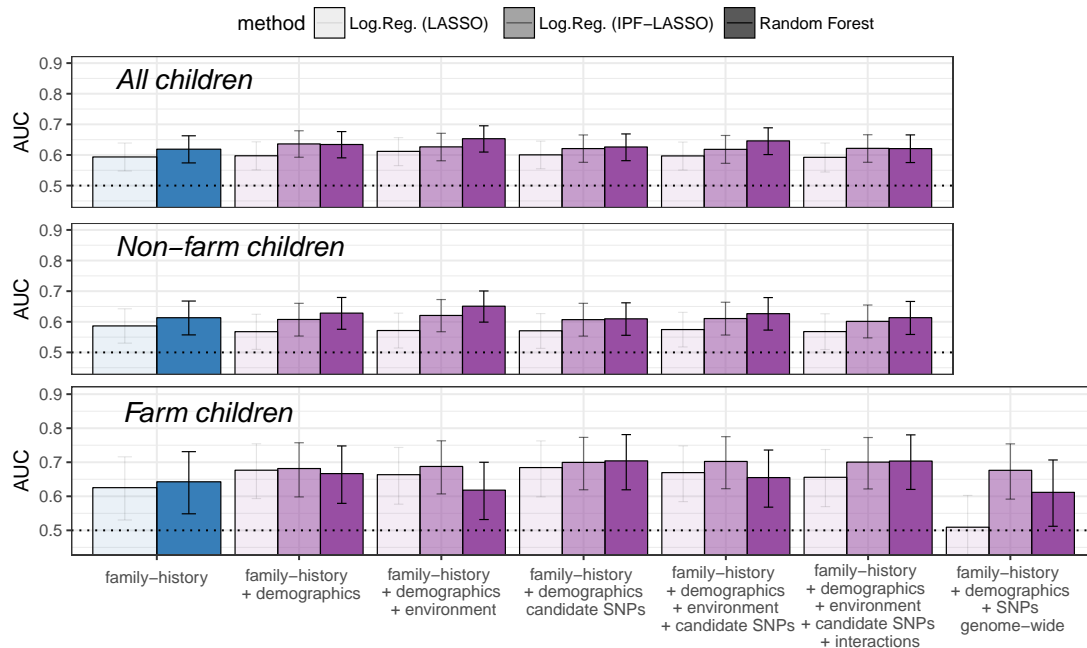


Figure 4.4: Prediction performance for family-history and additional predictors for different statistical methods. Prediction performance for models based on family-history alone and successively combined with demographics, environment or candidate SNPs, environment and candidate SNPs, environment interacting with candidate SNPs, and demographics plus genome-wide SNPs. As statistical methods, we used multivariable logistic regression with LASSO penalty, IPF-LASSO, and the random forest. The AUC is calculated as mean over 5 imputation data sets with 95% confidence intervals constructed by bootstrap using selection probabilities. The dotted line at 0.5 corresponds to the AUC-value where a prediction model classifies cases and controls not better than at random.

The most successful prediction models shown in Figure 4.4 were subsequently assessed for the contribution of individual variables to the entire model and externally validated. In the random forest prediction model for all children (AUC= 0.64 [0.54-0.73], Figure 4.5, left panel) and non-farm children (AUC= 0.63 [0.53-0.72], Figure 4.5, center panel) the variables for family history of asthma and atopy scored highest with respect to variable importance, followed by the demographics age and sex, but also 26 environmental exposure variables such as contact to cats, dogs, cows, straw, and hay. For farm children (Figure 4.5, right panel), also family history and sex contributed most importantly to the prediction models. Instead of environmental variables, however, three candidate SNPs emerged as significant predictors, one of them intergenic. The two other SNPs are known to be related to IL33 and RAD50 (Table 4.1). Sensitivity analyses using IPF-LASSO confirmed the IL33 SNPs from the random forest prediction model (Figure 4.6) with an AUC of

Table 4.3: Confidence intervals for bootstrap test of the difference between two AUCs for combined prediction model against a single determinant prediction model. For comparison of the AUC of the best prediction model combining several determinants (Figure 4.4) to the best prediction model for one determinant (always family-history, Figure 4.3), a one-sided bootstrap test for the difference between two AUCs was calculated with a level of significance of 0.05. The resulting lower bounds of the 95%-confidence interval of the differences in the AUCs are shown. For each subgroup (all children, non-farm children, farm children) the confidence interval for the difference to the AUCs does not overlap with the null — the AUC of the synergy model is significantly higher than the AUC for family history.

Subgroup	Best stand-alone model (Figure 4.3, random forest)	Best synergy model (Figure 4.4, random forest)	lower bound of 95%-CI for AUC difference
all children	family history	family history + demographics + environment	0.0051
non-farm children	family history	family history + demographics + environment	0.0078
farm children	family history	family history + demographics + candidate SNPs	0.0019

0.86 [0.59-0.99] averaged over the prediction scores of random forest and IPF-LASSO (see Supplemental Section A.2.4 for details).

A sensitivity analysis revealed AUCs of 0.57 [0.51-0.64] and 0.55 [0.51-0.58] for prediction by candidate SNPs and demographics in all children with and without a family history of asthma, respectively. External validation in the Austrian GABRIELA arm (Figure 4.7A) and the PASTURE birth cohort (Figure 4.7B, summary of variables see Supplemental Table A.2) confirmed the AUC values from the previously cross-validated random forest model. Sensitivity analyses using different asthma definitions yielded a better prediction quality for a model excluding individuals with current wheeze or asthma medication from the reference group (Figure 4.7C) and a model using only an asthma diagnosis irrespectively of obstructive bronchitis (Figure 4.7D).

Additional results for applying parametric IP bagging are shown in the supplement, Section A.2.5, which confirms that the approach works moderately for binary (environmental) variables but fails for SNP data which is neither symmetrically distributed nor binary as the scenarios in Chapter 3.

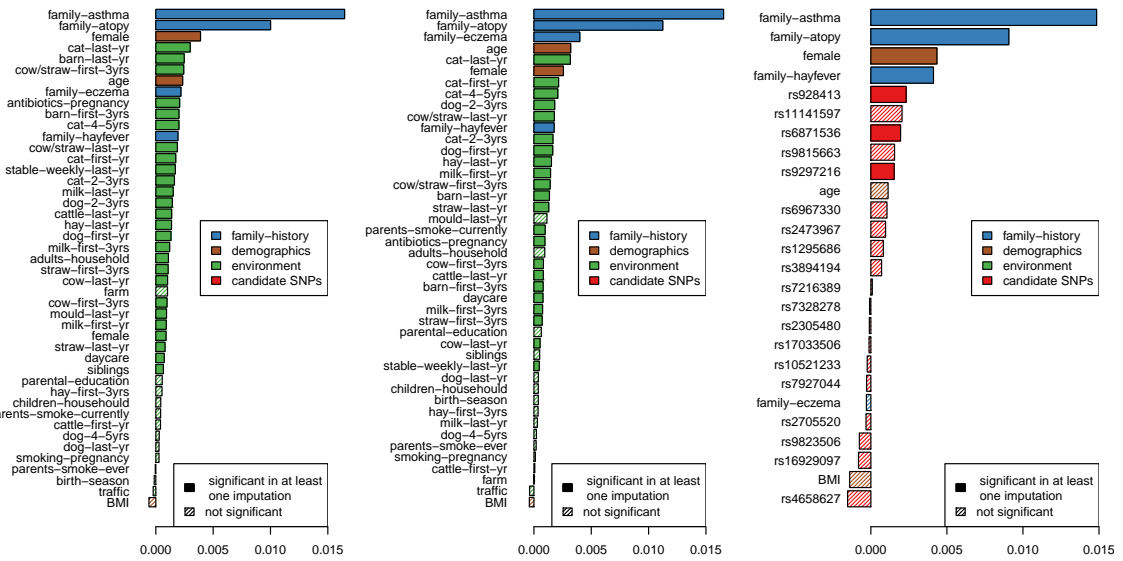


Figure 4.5: Importance of variables contributing to the best prediction models. Variable importance determined by random forest models for all children (left figure) and stratified for non-farm (center figure) and farm children (right figure).

4.5 Discussion

Asthma risk as predicted by logistic regression based on family history of asthma or atopy, demographics, environmental variables, and genetic data overlapped between cases and controls to a large extent thereby challenging disease prediction on an individual level. With the use of sophisticated methods from the area of machine learning, which allow for multivariable consideration of predictors, performance of prediction was improved noticeably beyond the classical logistic regression approach. In combined models, prediction of asthma was mainly driven by family history, sex, and various environmental variables, whereas candidate and genome-wide SNPs did not improve prediction. Only in farm children, genetic information contributed significantly to the prediction model, while environmental exposure did not add to prediction models in this group of children. Table 4.4 summarizes the top AUCs.

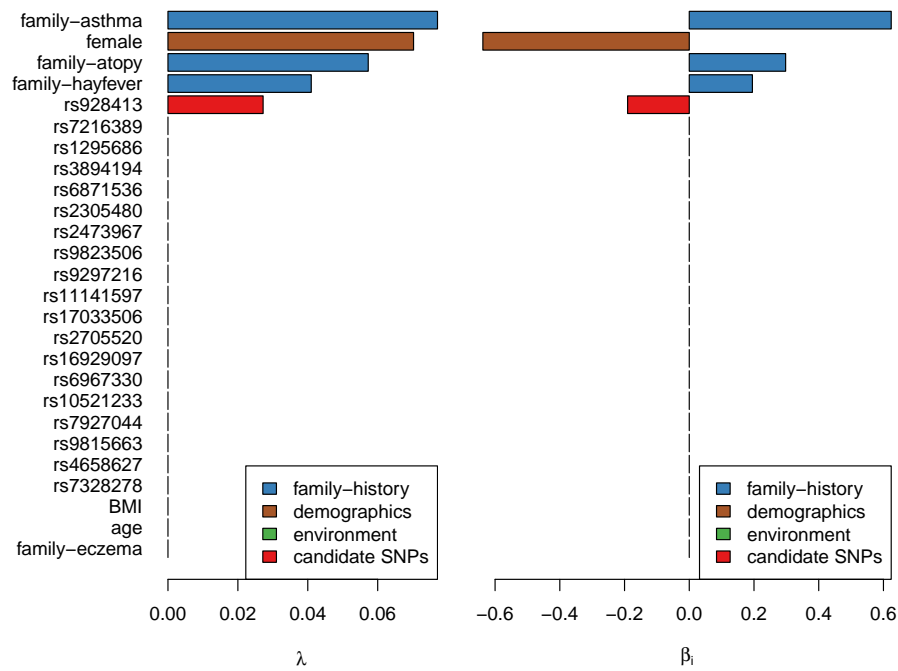


Figure 4.6: Variable importance by IPF-LASSO for the best model in farm children. Determinants of the prediction model IPF-LASSO in farm children which performed similarly to the random forest in the internal validation (Figure 4.4), sorted by importance (λ , left panel) with respective effect sizes (β_i , right panel). Positive β_i values represent risk factors; negative values represent protective factors; λ and β_i values are averaged over the 5 imputation datasets.

Unbiased validation guaranteed by incorporating sample bias and use of external data

Validation of prediction models was threefold: for a selection of a best model, cross-validation within the GABRIELA training data was performed; for a first external validation the Austrian GABRIELA arm and for a further external validation the PASTURE study were used. The first two validation procedures would be biased if performance was calculated without correction for sample selection bias. By correcting confidence intervals for the AUC via bootstrap using inverse-probabilities as selection probabilities we could guarantee that point and interval estimates were both unbiased. Further we provided a test for determining whether a prediction model performed significantly better than another one by again incorporating the principle of bootstrap with selection probabilities. By this, model comparison could be performed correctly.

Table 4.4: Top AUCs on different validation data. Best prediction was accomplished by analysis on farm children only. External validation using all children led to lower AUCs. In each case, the classifier was random forest.

Subgroup	Best AUC cross-validated (Figure 4.4)	Best AUC on Austrian GABRIELA arm (Figure 4.7 A)	Best AUC on PASTURE (Figure 4.7 B)
all children (family history + demographics + environment)	0.65	0.64	0.62
non-farm children (family history + demographics + environment)	0.65		
farm children (family history + demographics + candidate SNPs)	0.70		

Methodological shortcomings in previous prediction models overcome by multivariable modeling

The GWAS of the last two decades were definitively a success when considering the discovery of new loci and the confirmation or invalidation of candidate genes [129, 135]. However, the above-mentioned expectations with respect to disease prediction in individuals have not been fulfilled, regardless whether anticipated with hope or fear. Moffatt and colleagues already reported a low AUC of 0.58 for the seven top SNPs identified by their meta-analysis for childhood asthma [123]. However, this figure might be biased for the following reasons: First, the prediction model was fitted on the entire dataset leaving no independent sample for validation, which may have resulted in a too optimistic AUC. In our population such an approach would have resulted in an AUC of 0.60 for GWAS SNPs, instead of 0.54 as reported in Figure 4.3 (upper panel). Second, the approach chosen by Moffatt and colleagues integrated only the top seven SNPs thereby ignoring additional information provided by the SNPs that missed genome-wide significance or summary statistics in general, and thus may underestimate the true predictive power of a genome-wide approach [22, 135]. This second issue was exactly the starting point of our endeavor. We sought to integrate all available genetic information by multivariable modeling and to complement that with questionnaire data on familial predisposition and strong environmental determinants.

Therefore we applied random forest and various forms of penalized multivariable logistic regression such as LASSO, elastic net, and IPF-LASSO. These models find an optimal trade-off between statistical exhaustiveness and the risk of model overfitting; the latter might negatively impact external validity and thus predictive power.

For comparison, we also applied a two-step approach with first performing a classical simple logistic regression and second creating a prediction score based on the top hits of the previous step [191].

Analyzing modalities separately — best prediction by family history

Before combining environmental, genetic, familial, and demographic variables we assessed these four modalities separately (Figure 4.3). By our proposed solution for testing AUCs pairwise in a corrected manner, we showed that combining several of these compartments of variables leads to significantly better prediction than using them separately. With the use of random forest, we found significant prediction by environmental variables considering about 40 variables. Among those, exposure to cats and dogs contributed rather importantly to the prediction of asthma (Figure 4.5) though in GABRIELA these exposures were only related to atopic sensitization [88]. Otherwise, there were strong associations observed in this population with farm exposure reducing asthma risk by 37% and raw milk consumption by 45% [49], which contrast with the rather modest AUC-value of 0.55 [0.51-0.59] in the present analysis. Association estimates, however, are only meaningful in risk factor research or for prevention in populations, whereas prediction refers to individuals and requires much stronger associations with outcomes [69, 94]. Family history of asthma, e.g., increased asthma risk by 208% [49], which is reflected by the much better prediction achieved with the four variables concerning family history.

Random forest on genome-wide data as best prediction model purely based on genetics

Genetic effects, in turn, are rather weak in polygenic diseases. Even more the random forest prediction model by genome-wide SNPs with its AUC of 0.61 [0.51-0.70] in farm children is remarkable (Figure 4.3A, lower panel). It may reflect improved prediction by inclusion of SNPs above the genome-wide significance threshold. These non-significant SNPs might still be relevant for polygenic diseases and finally may help explaining the missing heritability [22, 135]. On the other hand, about 99.5% of the genome-wide SNPs did not significantly contribute to the prediction model and may have increased noise [169].

Combining modalities — SNPs overruled by family history?

When establishing combined prediction models based on several groups of variables, the genome-wide SNPs were replaced by candidate SNPs not among the top 50 genome-wide SNPs (Supplemental Figure A.7) and family history of asthma or other atopic diseases, which might be better proxies for predictive hereditary factors than the vast majority of genome-wide SNPs. Family history integrates a wealth of hereditary information though at much lower resolution as compared to genome-wide SNPs.

A family history may reflect shared environments such as the microbiome, which is clearly passed from mother to child [13]. Likewise a family history may represent conditions during pregnancy, e.g. an inflammatory status of the mother shaping the fetal immune system and thus contributing to disease transmission [47].

Nevertheless, family history might partially reflect the penetrance of the genotype, which is influenced by various biologic phenomena such as methylation and posttranslational modifications [96, 120, 170]. Consequently, SNPs that influence the phenotype might be detected more easily in a population enriched for a positive family history. This may also explain why the prediction model of diabetes type 1 susceptibility in a population preselected for a family history yielded a rather high AUC value of 0.87 [185]. However, this may only apply to rare diseases, whereas for highly common, heritable conditions such as asthma the predictive power of SNPs might be overruled by family history [41]. In the present analysis, prediction by candidate SNPs was weak and did not differ between the strata with and without family history.

Stratifying analyses for farm-exposure — highest predictive quality for farm children

Stratification for environmental exposure, however, was very informative: The prediction models between farm and non-farm children varied completely with respect to genetics. Farm children may be rather homogeneous in their environmental exposure leaving hardly any variance for the assessment of environmental factors thereby fostering prediction by genetic factors. Additionally, farm exposures may prevent many cases of asthma so that farm children might be affected mainly by genetically determined forms of asthma, which renders them an interesting population for genetic research.

Two of the SNPs contained in the random forest prediction model for farm children are

related to the genes IL33 (rs928413) and RAD50 (rs6871536) thereby representing two major asthma risk loci [18]. IL33 has been implied in allergies and autoimmune disorders, and a role in exuberant immune responses related to reduced numbers of regulatory T cells is discussed [155]. The other SNP was not selected in the IPF-LASSO probably because this approach gives more weight to other groups of variables as illustrated by the higher importance of sex (Figure 4.6). Nevertheless the detection of rs6871536 by random forest alone is informative. This SNP is situated in an intron of RAD50 in the TH2 cytokine locus on chromosome 5 and has been reported to be associated with asthma, atopic eczema, and total IgE levels [18, 109, 181].

Though marginally missing statistical significance, two further candidate SNPs (rs9815663, rs6967330) may also be of interest as they are related to CDHR3 and IL5RA. Like other members of the cadherin family of transmembrane proteins, CDHR3 is associated with asthma-related traits and a function in epithelial polarity, cell-cell interaction and differentiation has thus been suggested [18]. The alpha chain of the IL5 receptor is essential for differentiation and maturation of eosinophils, and inactivation of IL5 reduces airway eosinophilia [157]. Taken together, the detected genes are mainly related to the allergic aspects of asthma; allergic asthma, in turn, is related specifically to lung function impairment and need for inhaled corticosteroids [40]. In contrast, the SNPs of the asthma risk locus on chromosome 17q21 did not relevantly contribute to the prediction models in farm children. This locus has been suggested to encode susceptibility to environmental signals [32], which might not be relevant for the prediction of the sort of asthma farm children suffer from.

Improved prediction by omitting inconclusive asthmatics

Essentially, asthma is an umbrella term for various disease entities manifesting with similar symptoms [40, 45]. Children whose parents are not aware of an asthma diagnosis might be classified as controls even if they are treated with asthma drugs or experience current asthma symptoms. When excluding these children from the reference group (Figure 4.7), the prediction performed significantly better thereby implying true cases of asthma covered by this grey zone. The performance of the prediction also improved when asthma was defined irrespectively of recurrent diagnoses of obstructive bronchitis, which may point towards more severe asthma forms [40].

No improvement by including interaction terms

Previously it has been demonstrated that the GABRIELA study was well powered for detecting interactions with farming for SNPs with minor allele frequencies above 0.2 [48, 49]. Prediction, however, requires much stronger effects, which also includes larger interaction odds ratios. This might explain why inclusion of interaction terms with farming did not improve prediction quality in the entire population (Figure 4.4, upper panel). The association of farming with various SNPs on a genome-wide level was largely restricted to asthma cases (cf. Supplemental Figure A.7) thereby suggesting that farm children suffer from genetically distinct forms of asthma. However, assumptions about underlying pathomechanisms remain speculative as long as the farm effect on asthma has not been deciphered on a molecular level.

Fully exhausted methodology for prediction on high-dimensional multi-omics data

Technically we have fully exploited the instruments of predicting childhood asthma with modern machine learning methods and incorporated four sets of variables as efficiently as possible by optimizing their contribution potential via random forest and a multivariable penalized regression approach – the IPF-LASSO.

In general, multivariable techniques offer the opportunity to assess several predictor variables simultaneously thereby considering the complex correlation structure of multi-omics datasets and consequently the mutual interplay of variables. In addition, multivariable approaches reduce the risk of unstable statistical models, which may occur when relevant explanatory variables are missing.

Nevertheless multivariable regression still entails practical difficulties: In case the number of individuals is smaller than the number of predictor variables, the maximum likelihood estimator does not exist. This is particularly the case in GWAS and known as the $n \ll p$ problem. Also highly correlated covariates increase the variance of the parameter estimates and hence interfere with the stability of the estimated models. These difficulties can be overcome by penalized regression, particularly IPF-LASSO, which enables stable variable selection and simultaneously protects from overfitting by an optimized tuning parameter.

Consistently the highest prediction quality was achieved by random forest, a prominent representative of machine learning. In contrast to regression models it is based on decision trees and can efficiently handle high-dimensional data and incorporate interactions be-

tween predictor variables automatically. Random forest is unaffected by highly correlated variables and thus inherently robust.

Applied to genome-wide data, this combination of classifiers prevented from difficulties which can occur in this situation: As described in the introduction of this thesis, the optimal method for such data has to be established for each particular application as trait architecture, marker density and relatedness of SNPs vary from disease to disease [165]; with LASSO we employed the classical representative for a sparse model; with the conceptually completely different random forest we involved an appropriate pendant which contrarily can be seen as a dense model. Elastic net surely is closer related to the LASSO but forms a middle course between sparseness and density, as intrinsic variable selection is still performed but variables with weaker effect can remain in the model as shrunk effect. The additional application of the IPF-LASSO prevented other groups of variables from getting lost in the genome-wide data which contained at least tens of thousands times more variables than demographics, family history, or environment. Other approaches with different concepts have been proposed, for instance by incorporating Bayesian modeling [19, 128], but are partly not feasible for application for hundreds of thousands of variables or generally are less built for gaining best prediction power rather than for finding the correct associations.

Taken together we have applied computationally efficient, stable and robust methods, which run a low risk of model overfitting and can handle a high number of variables simultaneously and hence more appropriately. These properties render them ideal tools for prediction though they may be computationally demanding and require a powerful computing infrastructure. In addition we proposed and applied confidence intervals and a test of significance for comparing performances corrected for sample selection bias.

Concluding remarks

The insight that asthma runs in families is not trivial. As illustrated by our final models much of the predictive power of a family history was replaced by genetic and environmental variables – depending on the respective subpopulation. In other words, the simple question on a family history of asthma and atopy just integrates multifaceted information on several known environmental and genetic predictors and complements it with all the complexity of family life, which is neither captured by questionnaire records nor genome-wide data.

With this chapter we have used the knowledge we gained from the previous chapter, i.e. how to train prediction models when there is sample selection bias in the given data. As also validation data was biased, we developed solutions for incorporating appropriate correction strategies for this issue as well. Further, we extensively investigated the potential of predicting childhood asthma by environmental exposure and genetics. In the next chapter, we will extend the current project in several ways: we will integrate a further biological component for prediction, the immunology of an individual, thus have more modalities to integrate in a multi-omics situation, and, in this context, will treat a more complex case of missing data.

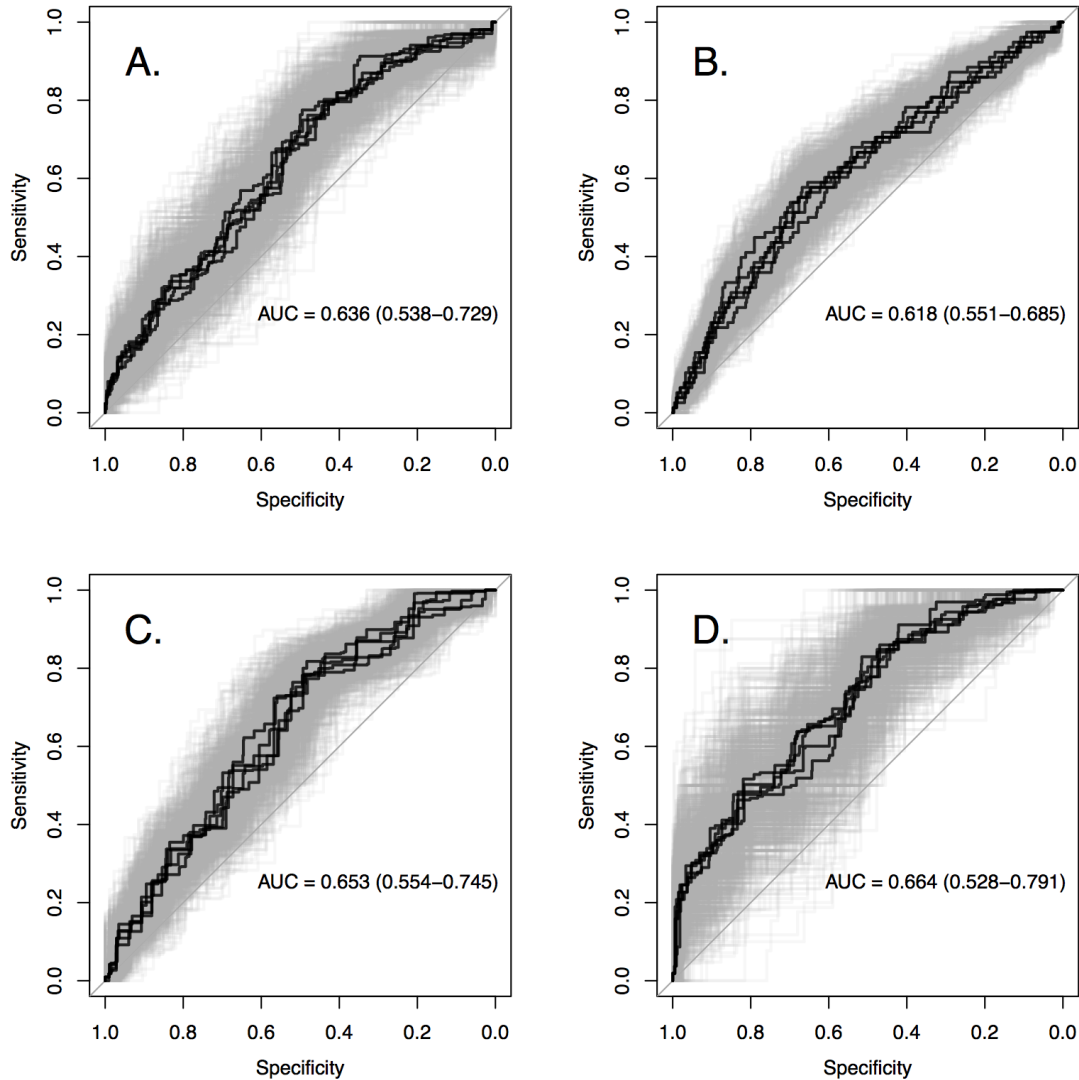


Figure 4.7: External validation of prediction models For external validation, individual ROC curves of the 5 imputations with AUC are shown. The internally validated random forest prediction model of a parent-reported doctor-diagnosis of asthma once or obstructive bronchitis twice based on family-history, demographics, and environment yielded a AUC of 0.65[0.61-0.70]. A. External validation in the Austrian arm of GABRIELA. B. External validation in the PASTURE birth cohort. C. Similarly to A, but excluding control individuals with current wheeze or use of asthma-specific drugs. D. Similarly to A, but for a parent-reported doctor-diagnosis of asthma once irrespectively of any obstructive bronchitis.

Chapter 5

Tackling multi-omics missingness patterns: classifying childhood asthma phenotypes using genetics, environment and immunology

Having gained knowledge on how well childhood asthma cases can be predicted by using questionnaires on environmental exposure and genetic SNP data, in this chapter, we will analyze data containing additional immunological components and will consider a modified definition of the disease.

As indicated previously, childhood asthma is not unambiguously defined. However, it can clinically be divided into two main phenotypes: allergic asthma (AA) and non-allergic asthma (NA) [145]. Several studies have tried to disentangle distinct underlying pathophysiological mechanisms, but were hampered by the complex nature of the disease [5, 29, 107, 139]. While singular targets were identified, one could not consistently pinpoint a reliable pattern of relevant pathways critical for asthma phenotype differentiation and in the long-term potentially patient-tailored treatment of the disease. However, this is important as a number of children with asthma are not well-controlled, potentially due to uniformal, not patient-specific therapies with mainly steroids to date.

Omics data, such as genomics and transcriptomics, have become increasingly available in human cohorts and thus more critical for understanding the pathogenesis of childhood asthma [177]. Inherent high dimensionality, incomplete data, and multiple platforms make

the analysis of prediction models complex. Reliable analysis strategies for multi-omics data from multiple platforms in large cross-sectional studies urgently needed to predict the risk of this multifaceted disease. Tools for integration of multiple omics datasets exist in literature [79] but are often restricted to analyzing correlation structures rather than building multivariable prediction models. Methods have been proposed to do so, i.e. using several modalities for predicting [176]. Acharjee et al. [1] use the machine learning method random forest and pre-select significant variables. Zhao et al. [195] analyze each modality on its own and merge the single components. The IPF-LASSO (see Section 2.3.4) incorporates each modality via penalized regression estimating weights for each modality. Approaches proposed in literature incorporate useful and efficient ideas. However, successful solutions for incorporating data structures where different values are measured for different observations are not yet available.

Hence, novel strategies to build and validate multivariable prediction models incorporating all individuals and all variables simultaneously are needed for classifying asthma in children. For cancer, patient-specific therapies are long-established, and this has been started for adult asthma by using biologicals such as anti-IL-5 for specific asthma phenotypes [127, 136]. For childhood asthma, patient-tailored therapies are still not available, but urgently needed to avoid long-term consequences of exacerbations.

In this study, we propose a novel approach to optimize prediction of childhood asthma phenotypes when different omics data types are used as input factors. Multi-omics data include questionnaire, clinical diagnostic, genotype, gene expression microarrays, quantitative real time RT-PCR (RT-qPCR), flow cytometry and cytokine secretion data. Combining multi-omics data types together with a novel and reliable analysis strategy for large human cohorts will contribute to detailed understanding of childhood asthma, potentially relevant for novel therapeutic strategies. The strategy can also be translated to numerous other complex diseases.

This chapter is in parts identical with the following manuscript:

[101]: Norbert Krautenbacher, Nicolai Flach, Andreas Böck, Kristina Laubhahn, Michael Laimighofer, Fabian J Theis, Donna P Ankerst, Christiane Fuchs, and Bianca Schaub. Classifying childhood asthma phenotypes from genetic, immunological and environmental factors: A strategy for high-dimensional multivariable analysis. *submitted*, 2018

5.1 Data collection

The following data collection methods have been performed by the clinical partners involved in this chapter's project or members who were involved in the study conduct.

5.1.1 Study population

Children between 4 and 15 years from southern Germany were recruited in the University Children's Hospital Munich from the CLARA/CLAUS (Clinical Asthma Research Association) study [139] in three groups, namely healthy children (HC), mild-to-moderate allergic asthmatics and non-allergic asthmatics. Parents completed a detailed questionnaire assessing health data on allergy, asthma, and socioeconomic factors. Asthmatic patients were diagnosed according to GINA guidelines [55]. Inclusion criteria for asthmatics were classical asthma symptoms, including at least three episodes of wheeze and/or a doctor's diagnosis and/or history of asthma medication in the past and lung function indicating significant reversible airflow obstruction according to American Thoracic Society (ATS)/European Respiratory Society (ERS) guidelines [16]. Allergy was defined based on a positive specific IgE level in accordance with clinical symptoms.

5.1.2 Modalities

We investigated seven data modalities: questionnaire, diagnostic, genotype, microarray, RT-qPCR, flow cytometry and cytokine data. Parents completed a detailed questionnaire assessing health data on allergy, asthma, and socioeconomic factors. Diagnostics included weight, height, blood count, immunoglobulins, CrP and IL-6 as well as FeNo.

5.1.3 Genotyping

Genomic DNA was extracted from whole blood (Flexigene DNA-Kit, Qiagen Hilden, Germany). Samples were genotyped for 101 loci using matrix-assisted laser desorption/ionization time-of-flight-mass-spectrometry (Sequenom, Inc., San Diego, CA). Deviations from Hardy-Weinberg equilibrium were assessed for quality control of genotyping procedures.

5.1.4 Microarrays

RNA of PBMC from a subgroup (14AA/8NA/14HC), comparable to the whole population, were analyzed by Affymetrix-GeneChip[®]Human-Gene 1.0 ST-arrays. Quality of scanned arrays was checked by MvA, density, RNA-degradation plots, using R and Bioconductor [85, 167]. Robust multichip averages were used for background correction, normalization, and control of technical variation.

5.1.5 RT-qPCR, flow cytometry and cytokines

Peripheral blood mononuclear cells (PBMCs) were isolated within 24h after blood withdrawal, cultured in X-Vivo (48h) unstimulated (U), stimulated with plate-bound anti-CD3 ($3\mu\text{g}/\text{ml}$) plus soluble anti-CD28 ($1\mu\text{g}/\text{ml}$), lipid A (LpA, $0.1\mu\text{g}/\text{ml}$) or peptidoglycan (PGN, $1\text{mg}/\text{ml}$, OR) at 37°C . Cell pellets were used for RNA isolation for microarray and RT-qPCR, and supernatants were used for cytokine measurements. RNA, isolated with RNeasy Mini-Kit was processed ($1\mu\text{g}$) with reverse transcriptase (Qiagen, Hilden, Germany). Gene-specific PCR-products were measured by CFX96 Touch[™] Real-time-PCR Detection-System (Bio-Rad, Munich, Germany) for 40 cycles. For flow cytometry, 2.5×10^6 cells were cultured in X-Vivo (48h, U, anti-CD3/28, LpA) and counted on a FACSCanto II flow cytometer (Becton Dickinson). Cytokine levels were determined in supernatants of PBMCs with Human Cytokine-Multiplex-Assay-Kit (Bio-Rad) using LUMINEX.

5.2 Computational and statistical analysis

Data Preprocessing

The statistical analyses were performed with R software [167]. We excluded non-asthma patients with other diseases so that only healthy children without any clinical allergic symptoms were included in the HC group. In addition, variables containing more than 25% missing values were completely removed from the data.

After restricting to the above listed phenotypes, our data set contained 260 observations. The variables can be sub-grouped into seven modalities according to their biological meaning: cytokines, SNPs, flow cytometry, diagnostics, questionnaire, and gene expression. Their dimensions and outcome distributions are summarized in Table 5.1.

Table 5.1: Dimensions and distributions of outcome (asthma) for the seven modalities.

Modality	#observations	#variables	HC	AA	NAA
Cytokines	148	39	74	63	11
SNPs	172	101	82	77	13
Flow cytometry	162	100	68	79	15
Diagnostics	248	24	103	117	28
Questionnaire	260	118	110	121	29
RT-qPCR	107	187	46	46	15
Gene expression	36	96953	14	14	8

The cytokine modality comprised 38 continuous variables of 148 observations of four different stimulation types. The genotype modality in terms of SNPs contained counts (0, 1 and 2) which were treated as continuous, i.e. we assumed the additive model (with 0 and 2 being the homozygous and 1 the heterozygous form). Thus 101 continuous variables of 172 observations were available. 100 physical and chemical characteristics of cells measured by flow cytometry were available for 162 subjects, all of them as continuous variables. The diagnostic modality contained results of detailed blood tests and individual characteristics. This modality contained 24 continuous variables and 248 observations. Questions about environmental exposure were answered and recorded in the questionnaire modality for 260 subjects with 118 questions. The gene expression modality contained polymerase chain reaction (PCR) measurements for 107 subjects and 187 variables. Data was available from 36 subjects for three stimulus types, resulting in a total of 96,953 variables for this modality.

For different modalities, diverse groups of children were observed, with different group sizes and only a small overlap between groups. This caused a complex missingness data structure between the modality data sets. In addition, there were missing values within in the modalities, which were assumed to be missing at random or missing completely at random.

Imputing missing values

The single data modalities originally containing numerous missing values were each completed via imputation yielding a basic structure of the full data set (Figure 5.1). We again used multiple imputation performed separately for each modality to handle missing values within each modality, described as follows. Again, first, we imputed missing values five times to generate several imputed (complete) data sets using the MICE algorithm. Next, we fit models described in the next sections to each of the complete imputed data sets, and finally, we averaged resulting estimates over the multiple data sets, and appropriated aggregated standard errors. We applied these multiple imputation steps to all analyses in this chapter according to Rubin [148] (as described in Section 2.2.3).

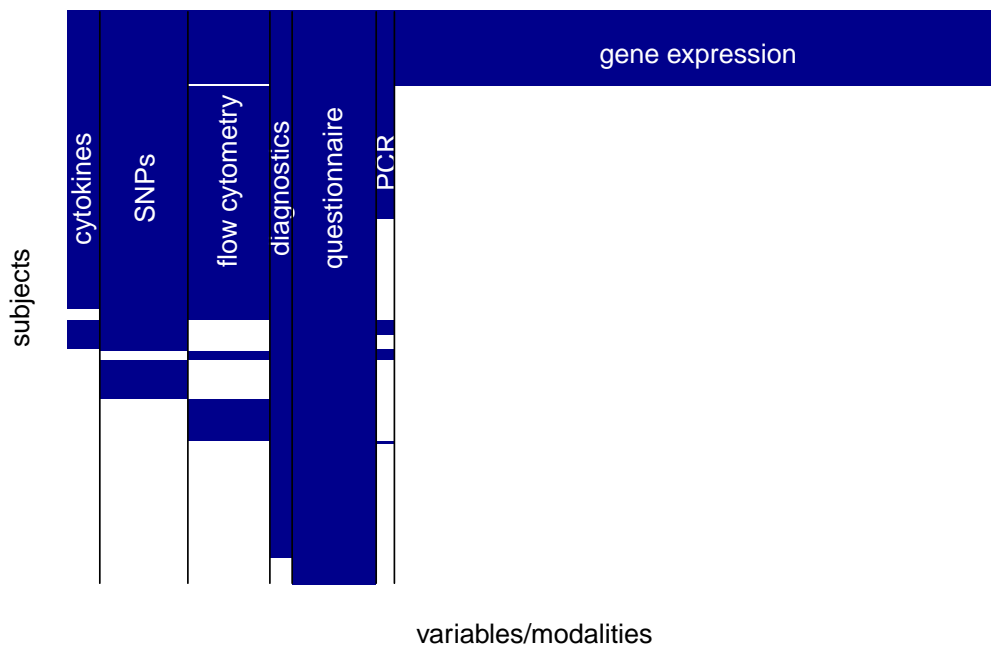


Figure 5.1: Structure of the given data after imputation within each modality. The blue-colored areas depict the given data values (all white areas correspond to missing data). The given data consists of seven groups of variables of the same type (modalities). There are only few subjects containing data for all modalities. The given gene expression by microarray data is the restricting component regarding complete cases and contains the most variables (reduced in figure for illustration reasons).

The intersection data sets after multiple imputation containing complete observations from all modalities embraced 33 children.

In the following, by $y_i \in \{0, 1, 2\}$ we denote the three-categorical outcome (coding “0” for representing category HC, “1” for category NA and “2” for AA, respectively) for individual i . In applications where generalized linear regression is involved, pairwise models with binary coding $y_i \in \{0, 1\}$ are used instead. We denote $m \in \{1, \dots, M\}$ as the modality, with $M = 7$ total number of modalities considered. All quantities referring to a certain modality are superscripted by this notation, i.e. x^m denotes the predictor matrix for modality m . A classifier trained on the data set (\mathbf{x}, y) and applied on a new individual with predictor values \mathbf{x}_{new} is then expressed as $\varphi((\mathbf{x}, y); \mathbf{x}_{new})$.

Validating predictions

To evaluate prediction quality of statistical models based on the given data, several classifiers were compared which were fitted on a training data set and validated on an independent test data set. Since only a small number of subjects were available, we implemented the more efficient leave-one-out cross-validation (LOOCV) approach, i.e. a prediction model was trained on $n - 1$ observations, using the left out observation for independent validation. By doing this for all n possible ways of partitioning the data in this LOOCV manner, we yielded exactly n independent predictions for validation.

We measured the loss of deviating predictions \hat{y} from the true values y by the ROC curve, in terms of the AUC and used the version for more than two outcome categories (cf. Chapter 2 Section 2.4), which we denote by $L_{AUC}(\hat{y}, y)$.

Applying statistical learning models

For predicting childhood asthma in terms of the three categories AA, NA and HC, we used statistical learning models employing regularization techniques that accommodate a greater number of variables than observations (cf. Section 2.3.3). Therefore, we used a regularized multinomial regression model as introduced in Section 2.3.2. We applied two versions of the penalized regression model; first, we used the LASSO penalty for dimension reduction as it performs hard thresholding by setting coefficients of non-predictive or strongly correlated variables to zero. Second, we used the elastic net penalty, so that some

coefficients of correlated variables are only shrunk towards zero, but without reaching the zero [76]. We also applied the two machine learning methods based on classification trees introduced in Sections 2.3.6 and 2.3.7 — random forest and stochastic gradient boosting.

Multi-omics learning approaches

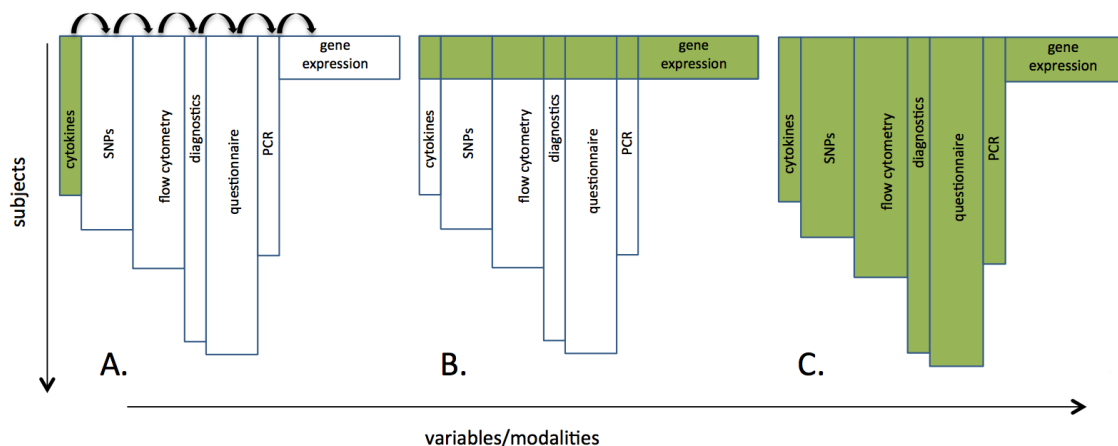


Figure 5.2: Schematic illustration of data partitions taken into account for prediction modeling at a time. A. All observations per modality were included, but training and validation was done separately for each block. B. Only complete observations were used, classifiers were trained on all modalities at once. C. All modalities and all observations were incorporated in a single prediction model and validated on complete observations.

We handled the complex data structures across modalities, i.e. having given many more values than the values for complete observations, by utilizing two different modeling strategies and developing a novel one from both ideas (see Figure 5.2) and compared all these in the scope of the four multivariable statistical learning approaches (multi-class logistic regression with LASSO penalty, multi-class logistic regression with elastic net penalty, random forest and boosting). In a first strategy (Strategy A), we analyzed each modality separately such that for each modality m all observations n^m were used but training and validation were possible only modality-wise (see Figure 5.3A). This approach could hypothetically yield a prediction model which performs better than using only few complete observations in case that there is one modality explaining most of the outcome because this modality would provide a sufficient number of observations. However, as this approach failed to combine the M modalities into one single prediction model, we jointly considered all modalities but took into account only complete observations as a second strategy (Strategy B). This yielded an overall prediction model but reduced the number

of subjects drastically from originally 260 subjects to 33 subjects. Figure 5.3B illustrates the strategy.

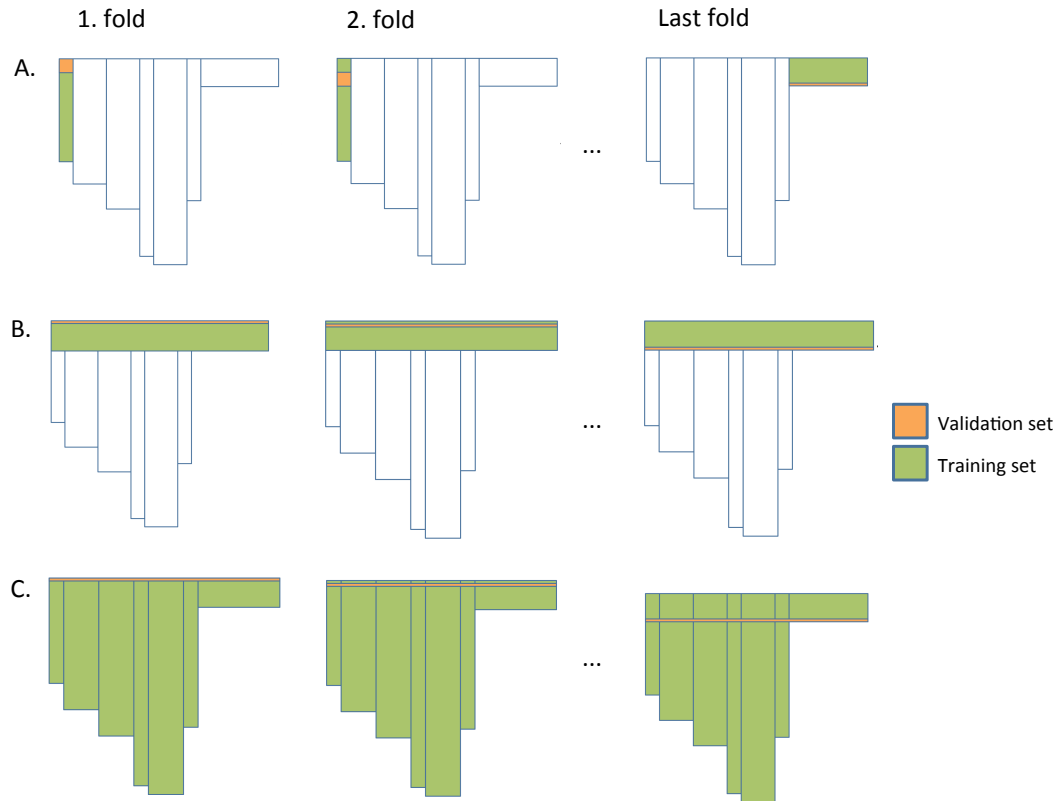


Figure 5.3: Three strategies of training classifiers on the complex missing data structure. Schematic boxes illustrate the data structure in terms of seven modalities partitioned into training (green) and validation (orange) sets. The columns per row illustrate the corresponding fold within a cross-validation procedure. A. All observations per modality were included; however, all classifiers were trained and validated separately for each modality (only depicted for the first modality). B. Only complete observations were used, classifiers were trained on all modalities at once. C. All modalities and all observations were incorporated in a single classification model and validated on complete observations.

Due to the shortcomings of the former two strategies, we developed a third approach incorporating all n observations and all M modalities (i.e. all variables p) in the next section.

5.3 A strategy for prediction on incomplete multi-omics data

We propose a novel model strategy shown in Algorithm 5.1 that makes use of all available data by learning a classifier on each modality separately and combining the single risk scores via optimized weights obtained in an inner-validation procedure depending on their single prediction power.

The strategy basically comprised two steps. After partitioning the data into training and test sets, in a first step, a classifier was fitted and validated on each modality separately. This enabled all observations per modality to be utilized for learning and prediction quality of each modality to be established. In the second step, the prediction quality (in terms of AUC) per modality served as weights for building a prediction model by combining the classifiers trained on the single modalities. The approach can be seen as a strategy of multi-view learning [192]. To implement the strategy we used LOOCV, with each of the 33 individuals with complete observations as new test observations and the remaining data for learning according to Algorithm 5.1. Figure 5.3C illustrates what the training-validation scenario on the given data looks like.

5.4 Results

260 individuals of the CLARA/CLAUS population with definitive phenotypes (AA/NA/HC) in total were available for the present analyses. AA cases (47%), NA cases (11%) and HC (43%) in the data differed with respect to variables from seven data modalities: cytokines, genetics, flow cytometry, diagnostics, environment, RT-qPCR and microarray (Figure 5.1, details Table 5.1).

5.4.1 Prediction modeling

Prediction of asthma risk with appropriate estimation of the prediction quality via cross-validation performed by two intuitive strategies (A and B) and a novel strategy combining both (Strategy C) yielded the following results. For preventing from severe overoptimistic bias regarding performance of a best model, we report results for all models [44, 92].

Strategy A performed prediction on single modalities separately. The comparison of performances on each modality on a stand-alone-basis showed no discriminatory power for

Algorithm 5.1: Weighted multi-omics prediction strategy

Input: Given data (\mathbf{x}, y) from modalities $m = 1$ to M ,
new observation \mathbf{x}_{new} , classifier φ

Output: Prediction score for new observation \hat{y}_{new}

1. Train the classifier separately on each modality:
For $m = 1$ to M

- (a) Partition data (\mathbf{x}, y) into training and testing sets denoted by $(\mathbf{x}, y)_{train}$ and $(\mathbf{x}, y)_{test}$
- (b) For each $\{(\mathbf{x}, y)_{train}, (\mathbf{x}, y)_{test}\}$
 - i. fit classifier φ on $(\mathbf{x}, y)_{train}^m$
 - ii. obtain prediction score on test observation: $\hat{y}_{test}^m = \varphi((\mathbf{x}^m, y)_{train}; \mathbf{x}_{test}^m)$
- (c) Let \hat{y}_*^m be the concatenation of the predictions \hat{y}_{test}^m for all test sets, and y_*^m be the concatenation of the corresponding observed values. Calculate the AUC $L_{AUC}^m = L_{AUC}(\hat{y}_*^m, y_*^m)$

2. Combine modality-wise prediction to one overall score:

- (a) Calculate weights w^m by

$$w^m := \frac{\mathbb{1}_{\{L_{AUC}^m > 0.5\}}(L_{AUC}^m - 0.5)}{\sum_{m=1}^M \mathbb{1}_{\{L_{AUC}^m > 0.5\}}(L_{AUC}^m - 0.5)}$$

- (b) Calculate prediction scores \hat{y}_{new}^m by

- i. fitting classifier φ on \mathbf{x}^m
- ii. obtaining prediction score on new observation: $\hat{y}_{new}^m = \varphi((\mathbf{x}, y)^m; \mathbf{x}_{new}^m)$

- (c) Obtain final prediction on new observation by

$$\hat{y}_{new} := \sum_{m=1}^M w^m \hat{y}_{new}^m$$

any classifier on flow cytometry (AUC for best classifier boosting 0.54[0.45-0.64]) and RT-qPCR (AUC for LASSO 0.47[0.36-0.59], Figure 5.4A). All CIs crossed the AUC=0.5 line, indicating that the prediction models did not do better than random guessing. There were moderate performances (mean AUC less than 0.7) for cytokines (boosting 0.60[0.51-0.70]), SNPs (random forest 0.66[0.57-0.75]), and diagnostics (LASSO 0.69[0.61-0.75]). Mean AUCs higher than 0.7 were yielded by modalities environment with an AUC for boosting of 0.75[0.69-0.82] and microarray with an AUC of 0.74 and a comparatively large confidence interval [0.54-0.90] (Figure 5.4A). Strategy B considered only observations with values of all modalities given and achieved a higher AUC than Strategy A for LASSO

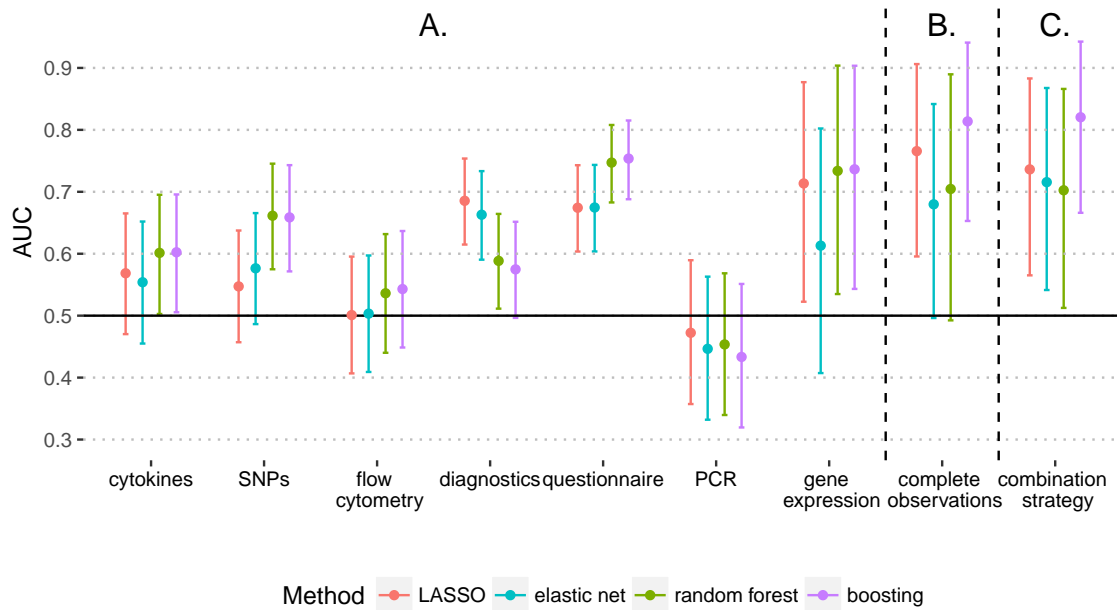


Figure 5.4: Comparison of prediction for different modalities for different statistical methods and strategies. A. Performance of prediction models on each modality analyzed separately (Strategy A). B. Performance for complete case model (Strategy B). C. Performance of combination strategy (Strategy C).

(0.77[0.60-91]) and boosting (0.81[0.65-0.94], Figure 5.4B), again with large confidence intervals. Strategy C combined A and B. Here, as in B, boosting outperformed the other classifiers clearly with an AUC of 0.82[0.66-0.94] (Figure 5.4C). Performance did not significantly increase from Strategy B to C. However, the classifiers' variance for C decreased slightly as shown by the narrower confidence intervals (Table 5.2). Thus, including not

Table 5.2: Length of confidence intervals for prediction models and Modeling Strategies B and C.

Classifier	CI length (Strategy B)	CI length (Strategy C)
LASSO	0.3106	0.3179
Elastic net	0.3455	0.3262
Random Forest	0.3973	0.3536
Boosting	0.288	0.2764

only all data modalities but also all observations per modality (Strategy C) may offer the chance to improve precision in risk estimates for asthma rather than it is possible by using e.g. only clinical or only diagnostic measures, or otherwise using all possible modalities but taking only those observations into account where all values for all these modalities

are measured.

5.4.2 Variable importance

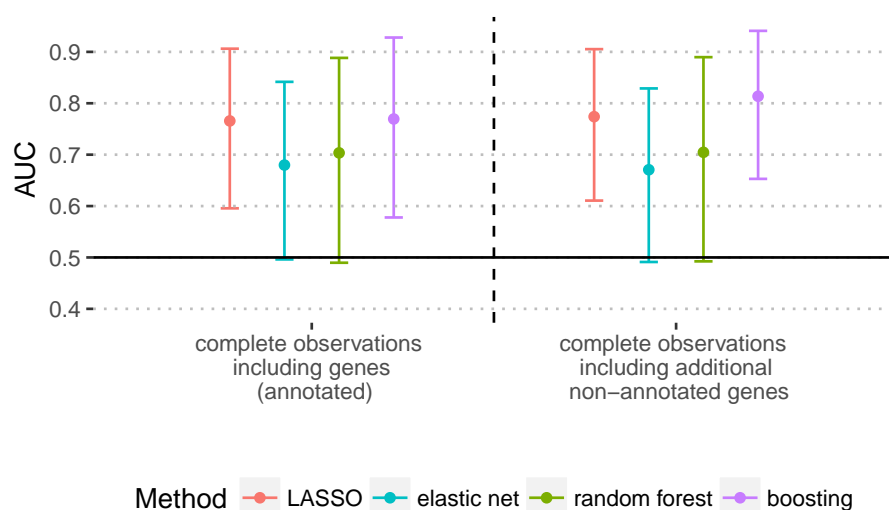


Figure 5.5: Performance of prediction models on the 33 complete cases (Strategy B). The procedure was run twice – once the modified model including genes which only contained annotated genes (left), once the original model including non-annotated genes in addition (right). The AUCs are calculated as the average over the 5 imputations; the error bars show 95% bootstrap confidence intervals.

Strategy B presents a reasonable trade-off between convenient interpretability and good prediction performance; contrary to Strategy C the entire prediction model consists of one model fit and not of a combination of several fits, so variable importance can be determined straightforwardly depending on the classifier. However, even though confidence intervals were larger, the performance of Strategy B was similar to the one of Strategy C. Hence, we investigated the best prediction model of Strategy B with respect to its most important predictor variables. For meaningful interpretation, we considered annotated genes only for the microarray modality set here. Figure 5.5 shows the performance of the refitted modified model, i.e. Strategy B with annotated genes only. Boosting, which originally performed best (AUC=0.81[0.65-0.94]), predicted slightly worse in the modified version (AUC=0.77[0.58-0.93]). Here, LASSO performed similarly to boosting (AUC=0.77[0.60-0.91]). Therefore, we analyzed the most important variables of both classifiers. As we based our investigations on variable importance on the two prediction models, we looked in detail at the ROC-curves for these two models (Figure 5.6); even though the overall

AUC was equal in both prediction models, their values differed regarding their one-versus-all comparisons (for boosting: AUC=0.79 for HC vs. all, AUC=0.78 for AA vs. all, AUC=0.72 for NA vs. all).

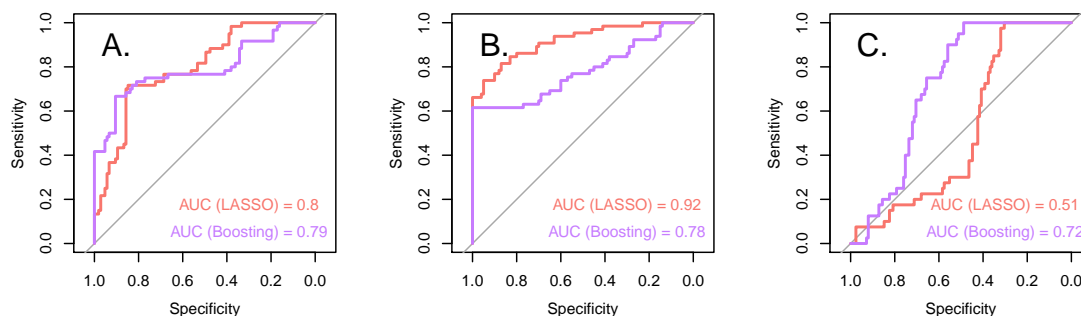


Figure 5.6: ROC-curves for the two best performing prediction models, LASSO and Boosting, on the 33 complete cases (Strategy B), when all variables were used but non-annotated genes were excluded. ROC-curves were calculated separately (aggregated over all 5 imputations) as A. Healthy controls (HC) vs. all others, B. Allergic asthmatics (AA) vs. all others and C. Non-allergic asthmatics (NAA) vs. all others. The overall AUC of 0.77 for both prediction models is a weighted average over the three single AUC-comparisons. The weights correspond to the proportions of HC (0.36), AA (0.39) and NAA (0.24), respectively.

Over all imputations, LASSO selected 22 non-correlated variables, which were exclusively genes from the microarray modality (Figure 5.7B and Table 5.3). In contrast, boosting used all variables by preferring and ranking them according to their importance without excluding correlated variables. Here, we took those 50 variables into consideration which were ranked highest. The selection contained variables from modalities microarray, cytokines, diagnostics, environment, and RT-qPCR (Figure 5.7A and Table 5.4). The two lists overlapped in three variables, illustrated by Figure 5.7C (and Tables 5.3 and 5.4), all of them were genes from the microarray modality: PKN2, PTK2, and ALPP. Thus, we considered these as model-independent most important variables for prediction of childhood asthma. A wider overlap could be determined with more relaxed assumptions (for details, see the supplement of this thesis, Section A.3.1), i.e. when variables in the two sets were considered as corresponding to each other when their correlation coefficient exceeded a pre-defined threshold.

In addition, we calculated importances of each modalities: we refit the originally best model (boosting with using also non-annotated genes again) repeatedly, each time leaving one modality out. Also here, the gene expression modality predominated (Figure 5.8).

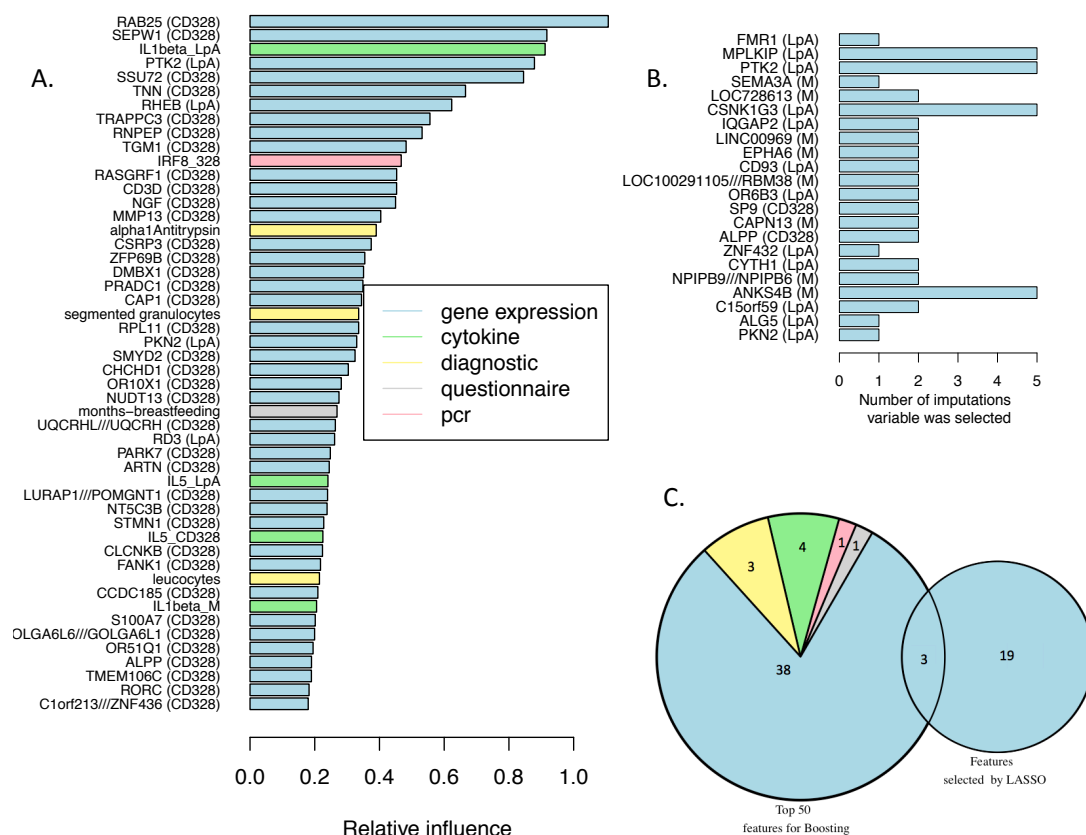


Figure 5.7: Variable Importance for best models on complete observations. Genes are denoted by their names with the type of stimulation in parentheses. A. Boosting variable importance: variables ranked under the top 50 by boosting in the complete case model averaged over all five imputations. B. LASSO-selected variables: variables selected by LASSO in the complete case model over all five imputations. C. Venn diagram/pie charts for sets of variables ranked highest by boosting (50 variables) and of variables selected by LASSO (19 variables). Three variables (genes) were selected in both prediction models.

5.5 Discussion

These are the first proposals for prediction analyses of childhood asthma using cytokine, genotype, flow cytometry, diagnostic, questionnaire, RT-PCR, and microarray data simultaneously which can be tested in further studies. Many studies on childhood asthma currently analyze phenotypes based on assessment of singular measurements e.g. of cytokines or gene expression data only [27].

Table 5.3: Variables selected by LASSO in the complete case model over all five imputations. Bold marked variables were also identified by boosting.

Gene	Occurrences in imputations	Description
PKN2 (LpA)	1/5	protein kinase N2
ALG5 (LpA)	1/5	ALG5; dolichyl-phosphate beta-glucosyltransferase
C15orf59 (LpA)	2/5	chromosome 15 open reading frame 59
ANKS4B (M)	5/5	ankyrin repeat and sterile alpha motif domain containing 4B
NPIPB9///NPIPB6 (M)	2/5	nuclear pore complex interacting protein family; member B9///nuclear pore complex interacting protein family; member B6
CYTH1 (LpA)	2/5	cytohesin 1
ZNF432 (LpA)	1/5	zinc finger protein 432
ALPP (CD328)	2/5	alkaline phosphatase; placental
CAPN13 (M)	2/5	calpain 13
SP9 (CD328)	2/5	Sp9 transcription factor
OR6B3 (LpA)	2/5	olfactory receptor; family 6; subfamily B; member 3
LOC100291105///RBM38 (M)	2/5	uncharacterized LOC100291105/// RNA binding motif protein 38
CD93 (LpA)	2/5	CD93 molecule
EPHA6 (M)	2/5	EPH receptor A6
LINC00969 (M)	2/5	long intergenic non-protein coding RNA 969
IQGAP2 (LpA)	2/5	IQ motif containing GTPase activating protein 2
CSNK1G3 (LpA)	5/5	casein kinase 1; gamma 3
LOC728613 (M)	2/5	programmed cell death 6 pseudogene
SEMA3A (M)	1/5	sema domain; immunoglobulin domain (Ig); short basic domain; secreted; (semaphorin) 3A
PTK2 (LpA)	5/5	protein tyrosine kinase 2
MPLKIP (LpA)	5/5	M-phase specific PLK1 interacting protein
FMR1 (LpA)	1/5	fragile X mental retardation 1

Combining several omics data types has optimized prediction of childhood asthma phenotypes in the CLARA childhood asthma study. The most important variables for prediction of childhood asthma phenotypes comprised novel identified genes, namely PKN2 (protein kinase N2), PTK2 (protein tyrosine kinase 2), and ALPP (alkaline phosphatase placental).

In addition to the complexity of modeling seven groups of variables of various dimensions, which we called modalities, a further challenge required a novel strategy: Complete observations, i.e. observations where values were given for all modalities, occurred only rarely; for subgroups of modalities, many more observations were available. The novel strategy had to incorporate all individuals and all variables at the same time with respect to building and validating multivariable prediction models. We solved the missing data

Table 5.4: Variables ranked under the top 50 by boosting in the complete case model averaged over all five imputations. Bold marked variables were also identified by LASSO.

Variable	Importance	Modality	Description
C1orf213//ZNF436 (CD328)	1.11	array	chromosome 1 open reading frame 213// zinc finger protein 436
RORC (CD328)	0.92	array	RAR-related orphan receptor C
TMEM106C (CD328)	0.91	array	transmembrane protein 106C
ALPP (CD328)	0.88	array	alkaline phosphatase; placental
OR51Q1 (CD328)	0.85	array	olfactory receptor; family 51; subfamily Q; member 1
GOLGA6L6//GOLGA6L1 (CD328)	0.67	array	golgin A6 family-like 6// golgin A6 family-like 1
S100A7 (CD328)	0.62	array	S100 calcium binding protein A7
IL1beta_M	0.56	cytokine	IL1beta_M
CCDC185 (CD328)	0.53	array	coiled-coil domain containing 185
Leucocytes	0.48	diagnostic	Leucocytes
FANK1 (CD328)	0.47	array	fibronectin type III and ankyrin repeat domains 1
CLCNKB (CD328)	0.45	array	chloride channel; voltage-sensitive Kb
IL5_CD328	0.45	cytokine	IL5_CD328
STMN1 (CD328)	0.45	array	stathmin 1
NT5C3B (CD328)	0.40	array	5'-nucleotidase; cytosolic IIIB
LURAP1//POMGNT1 (CD328)	0.39	array	leucine rich adaptor protein 1// protein O-linked mannose N-acetylglucosaminyltransferase 1 (beta 1;2-)
IL5_LpA	0.37	cytokine	IL5_LpA
ARTN (CD328)	0.35	array	artemin
PARK7 (CD328)	0.35	array	parkinson protein 7
RD3 (LpA)	0.35	array	retinal degeneration 3
UQCRHL//UQCRH (CD328)	0.34	array	ubiquinol-cytochrome c reductase hinge protein-like// ubiquinol-cytochrome c reductase hinge protein
months-breastfeeding	0.34	questionnaire	months-breastfeeding
NUDT13 (CD328)	0.34	array	nudix (nucleoside diphosphate linked moiety X)-type motif 13
OR10X1 (CD328)	0.33	array	olfactory receptor; family 10; subfamily X; member 1
CHCHD1 (CD328)	0.32	array	coiled-coil-helix-coiled-coil-helix domain containing 1
SMYD2 (CD328)	0.30	array	SET and MYND domain containing 2
PKN2 (LpA)	0.28	array	protein kinase N2
RPL11 (CD328)	0.27	array	ribosomal protein L11
Segmented granulocytes	0.27	diagnostic	Segmented granulocytes
CAP1 (CD328)	0.26	array	CAP; adenylate cyclase-associated protein 1 (yeast)
PRADC1 (CD328)	0.26	array	protease-associated domain containing 1
DMBX1 (CD328)	0.25	array	diencephalon/mesencephalon homeobox 1
ZFP69B (CD328)	0.24	array	ZFP69 zinc finger protein B
CSRP3 (CD328)	0.24	array	cysteine and glycine-rich protein 3 (cardiac LIM protein)
alpha1Antitrypsin	0.24	diagnostic	alpha1Antitrypsin
MMP13 (CD328)	0.24	array	matrix metalloproteinase 13 (collagenase 3)
NGF (CD328)	0.23	array	nerve growth factor (beta polypeptide)
CD3D (CD328)	0.23	array	CD3d molecule; delta (CD3-TCR complex)
RASGRF1 (CD328)	0.22	array	Ras protein-specific guanine nucleotide-releasing factor 1
IRF8_328	0.22	pcr	IRF8_328
TGM1 (CD328)	0.21	array	transglutaminase 1
RNPEP (CD328)	0.21	array	arginyl aminopeptidase (aminopeptidase B)
TRAPPC3 (CD328)	0.21	array	trafficking protein particle complex 3
RHEB (LpA)	0.20	array	Ras homolog enriched in brain
TNN (CD328)	0.20	array	tenascin N
SSU72 (CD328)	0.19	array	SSU72 RNA polymerase II CTD phosphatase homolog (S. cerevisiae)
PTK2 (LpA)	0.19	array	protein tyrosine kinase 2
IL1beta_LpA	0.19	cytokine	IL1beta_LpA
SEPW1 (CD328)	0.18	array	selenoprotein W; 1
RAB25 (CD328)	0.18	array	RAB25; member RAS oncogene family

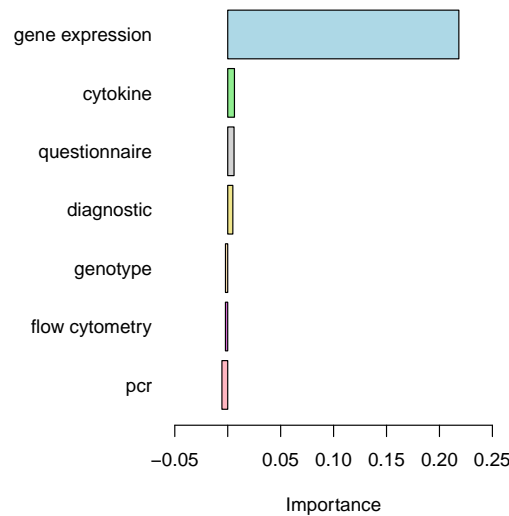


Figure 5.8: Importance of modalities. The importance measure is the difference of AUCs of the full model (AUC=0.81) and the reduced model. For all models boosting was used and averaged over the 5 imputations. The microarray modality is clearly the most important modality.

issue within each modality by multiple imputation since missingness completely at random could be assumed for this type of missingness [173]. After filling the data gaps within each modality, only a minority of 33 observations contained values for all modalities. We considered and compared three modeling strategies combined with four modern statistical learning methods, LASSO, elastic net, random forest, boosting. All four classifiers are capable of handling biomedical data difficulties, such as highly-correlated variables and large numbers of variables, partly exceeding the number of observations.

Prediction by seven modalities — best prediction obtained by using boosting

The first intuitive modeling approach (Strategy A), which trained and validated prediction models on each modality separately, showed differences in prediction quality in both, the different modalities and the four different classifiers. Prediction was unambiguously successful for environment and microarray, the only modalities with all prediction models performing significantly better than random guessing (Figure 5.4A). According to this criterion, prediction using cytokines, genetics, or diagnostics was successful, however, given that a certain classification algorithm was selected. Flow cytometry and RT-qPCR modal-

ities alone showed no evidence of predictive power, irrespectively of which classifier was used. This is important as a number of studies are analyzed based on singular techniques associated with childhood asthma phenotypes.

A further modeling approach (Strategy B), using only complete observations for prediction, proved by increasing performances that combining all variables of all modalities to one model is more predictive than using only single modalities.

Both strategies were trade-offs between using all observations per modality for the fitting process and using all modalities simultaneously in a single prediction model. Combining both aspects for training a prediction approach for the underlying complex missing data structure required a special strategy. Therefore, we developed a novel modeling approach (Strategy C), using the complete data for the training process (Figure 5.2C) by training a classifier and optimizing a weight via internal model validation for each modality separately in a first step and aggregating all established components in a second step (Algorithm 5.1). This strategy decreased the variability of asthma prediction on independent data (Table 5.2). Boosting showed best performance for both, the complete cases model and our novel strategy (Figure 5.2B and C). This method is helpful for clinical data sets where a multitude of immune-related measurements are available but missing or small numbers of subjects pose a problem for common analysis strategies.

Note that in this project the best model's AUC value should not be recorded as generalizing performance on new data since no external validation data was available and due to the extreme limitation of sample size any further partitioning into a more nested resampling procedure was not feasible (cf. Section 2.4.2).

Contributinal influences — gene expression is most predictive

Prediction on complete cases using annotated genes only was comparable to the original model with using also non-annotated genes and yielded high interpretability regarding the most important variables for prediction. We thus repeated prediction by Strategy B on the adjusted selection of genes. Evaluation by two conceptually different methods, the variable selection via LASSO and the relative influence determined by decision trees in the framework of boosting, yielded three model-independent most important variables for prediction: the genes PTK2, PKN2 and ALPP.

PTK2, a member of focal adhesion kinase (FAK), encodes a cytoplasmic protein tyrosine kinase that localizes to focal adhesions and contributes to integrin-mediated cell processes related to cell survival. The activation of this gene regulates a wide variety of cellular

responses and is assumed to be important in the early step of cell growth and intracellular signal transduction pathways [134]. Although tyrosine kinases play an important role in several pulmonary mechanisms like in airway hyperresponsiveness and airway remodeling, no correlation between PTK2 gene and asthma has been described so far [72]. PKN2, also called protein kinase C-related kinase 2 (PRK2) is a Rho target protein which regulates the apical junction formation in human bronchial epithelium. It has been shown critical for human cancer, and would represent a novel gene pathway potentially relevant for childhood asthma [180]. ALPP is a gene which encodes the placental alkaline phosphatase that catalyzes the hydrolysis of phosphoric acid monoesters and was previously identified to be potentially involved in recurrent spontaneous abortion [175]. Although these three genes have not been associated with childhood asthma yet, the findings in this study could be a first hint for future investigations.

Further model-specific variables contributing to prediction were obtained (Table 5.3 and Table 5.4). Contrary to the LASSO model which only labeled genes as most important, boosting found variables also from other modalities. One of them is the number of months of breastfeeding. This may have an influence on asthma, however, can be a case of translucent correlation since mothers with family history may be biased in their decisions for breastfeeding. Besides this, selected cytokines such as IL-1 β and IL-5, diagnostics variables and RT-qPCR variables such as IRF8 have been identified as important by boosting (Figure 5.7A).

In our results no genotype variables (SNPs) turned out to be important for prediction. This is not surprising as in our and in previous analyses SNPs on stand-alone basis did not exceed AUC values of around 0.60 [123]. The low predictive effects of SNPs may be covered by effects from other omics data sets in our analysis.

Prediction techniques — using well-established algorithms and all data information

We have used four of the most powerful instruments for prediction in terms of classification from regularization regression methodology to machine learning. In practice, classical approaches as (multivariable) non-penalized logistic regression can bias parameter estimates and make models instable when variables are highly correlated. Further, there is no maximum likelihood estimator when the number of variables exceeds the number of observations. Particularly the microarray data set represents both difficulties. Penalized regression, such as the used LASSO and elastic net, solve these problems; variable

selection generally ensures stability and prevents from overfitting.

Conceptually different but equally sufficient prediction methods are ensembles of decision trees, commonly random forest and boosting, as used in our analyses. Both belong to the most popular methods in machine learning and are now used in immune-related analysis. They can handle highly-correlated variables and high-dimensional data as well and incorporate interactions between contributing variables. The ensembling principle combines many decision trees at once and thus makes the two methods highly robust.

Apart from applying efficient classification algorithms, and running and comparing three modeling strategies overcoming the complex data structure completed the methodology of predicting childhood asthma: multi-omics approaches for childhood asthma have been proposed [56] but rather for finding associations than for building multivariable prediction models. Predicting on each modality separately revealed first answers on the predictive power of each modality when the full information given was used. However, this did not incorporate the multivariate structure between the modalities and could hence cause an information loss. The obvious solution to only use complete observations with respect to all modalities, again, came at the cost of a lack of information due to a smaller number of observations. Prediction seemed complete and fully efficient only if all variables and all observations were included in the analysis. Our novel approach, combining weighted prediction scores obtained from the full information of each modality, fulfilled this requirement.

In conclusion, we applied robust and stable classification algorithms in concordance with strategies for fully exploiting all information of the data in order to yield best possible results for predicting asthma from seven modalities from genetics, immunology and environment. Penalized regression methods complemented with machine learning approaches have not been used on several modalities for prediction of childhood allergic asthma and non-allergic asthma so far and should be considered as efficient prediction methods for this kind of application and beyond. Prediction analysis on incomplete data with respect to different modalities is feasible with certain strategies. We developed a novel strategy combining all information from the data leading to smaller prediction variability. However, the sufficient performance of the complete-case prediction model suggests focusing future data collection on enriching complete observations rather than enlarging the number of investigated individuals in total. This is important and requires a strict and thorough recruiting protocol, which is particularly difficult in children and if multicenter studies are envisioned. Microarray data in terms of three target genes responsible for integrin-mediated cell processes, regulation of apical junction formation in human bronchial epithelium and

placental alkaline phosphatase are predictive for asthma independently of the model approach, even though model-specific results show contributions from other modalities, such as breastfeeding-months, IL1-beta and IL-5 cytokine and IRF-8 gene expression.

For the future, we suggest to implement our novel analysis strategy to more comprehensively understand and analyze complex human immune regulation with respect to childhood asthma phenotypes. This method is also applicable for other cohort studies aiming to assess multi-omics data sets in medium or large cohort studies. Further, when more data like in the given study can be made available there is high potential for building and improving current risk tools for childhood asthma which can be optimized by distinguishing for pairs of outcome categories as Ankerst et al. [8], for instance.

Chapter 6

Summary and perspectives

6.1 Summary

This thesis provides statistical methodology in the context of predicting disease risk on complex data.

The main methodological contribution is given by Chapter 3: We treated the problem of having data sets under sample selection bias at hand and investigated the problem of how to correct for this bias when the data arises from a stratified random selection process or from one- or two-phase case-control studies. Even though there are approaches from literature for correcting general classical statistical analyses there was a gap for the area of machine learning. Correction methods for these have been provided in literature or could be modified from similar problems to the one at hand but our simulation study confirmed that these are not satisfactory in several scenarios. Therefore we proposed two new methods where one of them — the parametric inverse-probability bagging — could solve the issue and in all scenarios outperform those prediction models which were not corrected. Especially the case of the random forest as prediction model showed no satisfactory results for any correction method but for the parametric inverse-probability bagging. For other classifiers our approach performed at least as good as other correction approaches. Taken this together, by Chapter 3 we showed how approaches from similar fields can be adjusted or modified to be suitable to the field of sample selection bias in stratified random samples. We compared these and other state-of-the-art approaches and showed in which scenarios and for which prediction models which of the approaches are eligible for correcting sample selection bias appropriately. We finally provided novel methodology that fills those gaps where all other approaches fail and implemented the

relevant software in terms of an R package which is publicly available.

Theory treated in Chapter 3 was an essential basis for the following project: By Chapter 4 we showed an application for the issue of sample selection bias on real data for the example of predicting childhood asthma from genetic and environmental variables. We leaned on results from Chapter 3 to adjust appropriately for the bias coming from a two-phase stratified random selection process and provided a strategy on how to find a powerful prediction model that performed well on new data. We did this in the more difficult context of having extremely high-dimensional data in terms of an n -smaller- p problem where p was in the range of hundreds of thousands to millions of variables. For appropriate validation of different models within the given data we proposed methodology based on bootstrap: it provides weighted model validation with additional confidence intervals and a test of significance in order to estimate uncertainty of predictions on the one hand and to test if a prediction model performs significantly better than another one on the other hand; all this is done in a weighted manner, so that correction for sample selection bias is guaranteed. In addition to Chapter 3 where *learning* under sample selection bias was investigated, this methodological contribution treated sample selection bias, but for *validating* on new data. From the biological point of view we showed that family history of asthma and atopy remains an important variable for prediction as it represents complexity of family life which cannot only be covered by using questionnaire data on environmental exposures or genome-wide data. Thus, the assumption that genome-wide data can predict polygenic diseases might have been exaggerated — at least for childhood asthma.

By Chapter 4 the foundation was laid for an even more complicated situation; with a further data set on childhood asthma containing again genetic and environmental variables but this time with diverse immunological variables in addition, we further contributed to methodological development in the area of prediction of disease risk. We did so especially for another complex issue in data stemming from a clinical study: data for several different modalities was given but for each modality different observations contained measured values and only few complete observations existed. As also high-dimensionality was an issue here we incorporated several statistical and machine learning methods into strategies of handling the complex missing data issue. We finally provided a strategy for learning an overall prediction model that incorporated both, all observations and all variables, so that all given information in the data was used. The proposed strategy predicted asthma risk successfully and increased precision of risk estimation compared to other strategies.

In addition to novel methodology, the application to asthma in the Chapters 4 and 5 showed important novel implications in medicine: Both projects were investigations of

how well childhood asthma phenotypes can be predicted. In the first one (Chapter 4) data on farm-related environmental exposures and on the genome were available in addition to family history and demographic variables. Besides determining the prediction accuracy the goal was to identify which predictors are most relevant. We identified family history, a determinant integrating environmental exposure and genotype, as the most relevant predictor. Prediction accuracy was only moderate but showed an improvement for investigating only farm children. These are generally strongly protected through environmental exposure, but showed only genetic polymorphisms as predictive variables rather than environmental factors.

Contrary to this study, only a fraction of patients was available in the second asthma project (Chapter 5). The goal of distinguishing healthy children, mild-to-moderate allergic asthmatics and non-allergic asthmatics was investigated without external validation, but yielded valuable discoveries: using immunological variables in addition to environmental and genetic variables showed good performance, especially because of using gene expression data which contained the three most important predictors that have not been associated to childhood asthma before.

6.2 Outlook

In the following we will provide some perspectives for the methodology we have dealt with in this thesis. We will give several examples how the content of this thesis can be used and further developed in future research.

Correction methods for class imbalance problem

In the project on sample selection bias theory (Chapter 3) we investigated the situation where categories of certain variables are rare in the population but get enriched when a sample is taken. That is, the affected variable was balanced out in a way that its categories occurred with equal frequency. This sample process caused sample selection bias in the resulting given data set. Approaches were utilized which in diverse ways oversampled the originally more frequently occurring category so that the original data distribution was generated.

This situation where adjusting for the distribution of the population is the goal may seem to be the common case in which oversampling usually comes into play. However, there

are further situations where oversampling techniques are applied but in a kind of opposite manner: in so-called class imbalance problems [2], a variable which is typically the outcome variable contains a rare category in the *given* data set. Here, regardless of the original distribution in the population, oversampling techniques are applied in order to achieve that classifiers can be trained on data with a more balanced outcome variable [91, 105, 108]. This is done because several classifiers generally fail to learn successfully on data with imbalanced outcome variables. Thus, contrary to the sample selection bias situation, here, given data is manipulated the other way round: a rare category is oversampled to make the variable balanced rather than a balanced variable is oversampled to generate an original distribution of the population.

One successful approach for class imbalance which generates synthetic observations for this problem is SMOTE [34], for instance. In Chapter 3 we had modified this approach to use it for the purpose of overcoming the sample selection bias problem. Even though it was not successful in this case, the pendant could be successful: Our two approaches, the stochastic inverse-probability oversampling or the parametric inverse-probability bagging, which originally are designed for correcting sample selection bias also generate synthetic observations and could be applied for class imbalance problems instead. They could be used for oversampling a rare category in the data at hand. This may outperform other approaches applied for the class imbalance issue, especially when a parametric generation of observations is suitable for the data.

For realizing the idea of modifying our proposals into class imbalance solutions, further refinements have to be investigated extensively; complete oversampling of the rare category may not be the most ideal procedure: for instance, undersampling of the majority category or a combination of under- and oversampling majority and minority categories, respectively [34], may be required to make the approach most effective.

Sample bias incorporation in other fields

By Chapter 3 we showed a comprehensive investigation of how to act in statistical analyses when a sample is not taken at random but observations have different probabilities of being included in the sample. Then the given data suffers from sample selection bias. This kind of situation does not only occur in epidemiological studies. Another highly relevant example is single cell RNA sequencing (scRNAseq) analysis: scRNAseq is now a high-throughput method which provides gene expression profiles, i.e. measures gene activities, of cells whose number is quantified in addition [78]. By scRNAseq techniques whole tissue samples can be characterized. A sample in this context is defined as the isolated single cells that are

measured at once in a single run. When such a tissue sample from an organ is investigated, many different cell types can be identified. A question of interest is how the abundances of cell types change when a certain disease occurs or is present. In order to investigate rare cell types in this context also here enrichment strategies within the cell sorting (flow cytometry [14]) can be applied to increase statistical power for detecting effects. Usually correction methods are not common in this field, and thus could be integrated. Depending on the type of analysis either classical correction methods (as for logistic regression) or similar methodology as proposed and compared in Chapter 3 based on bootstrap or bagging can be directly applied or extended to guarantee unbiased statistical inference or successful prediction in this new field.

Improved asthma risk score based on multi-omics data

In this thesis we have investigated how well childhood asthma can be predicted in terms of how well asthma cases are distinguishable from healthy controls. However, clinicians often would like to achieve something more: ideally a personalized risk score can be provided, directly applicable to a child for which it is unclear whether asthma will evolve or not and telling the probability of getting the disease. Such risk scores have been provided: the asthma predictive index [33] is a widely used risk score tool for childhood asthma [86] which is based on predictive variables like family history, diagnosis of eczema, sensitivity to allergens, food allergy and wheezing symptoms. However, it does not include genetic or immunological factors.

Chapter 4 has approved that family history is indeed the most important predictive variable, but for the case of investigating only farm children also genetics contributed further to improve the prediction of asthma risk. Moreover, Chapter 5 showed high potential for risk prediction when immunological components are incorporated.

Thus, there is a high chance that an individual risk score for childhood asthma can be improved by incorporating genetic and immunological factors. Therefore more multi-omics data sets on childhood asthma including these diverse groups of variables with a sufficient number of observations could lead to a wider selection of important influence factors from various data types by applying modern statistical learning methods. This could eventually result in the successful development of an improved personalized risk score for childhood asthma.

Deep learning involving features measured over time

“Big data” has gotten a famous buzzword in the last years and has been interpreted in many different ways. One intuitive interpretation surely is big data in terms of large data sets. In this work we have had large data sets with respect to the number of variables p . However, big data in the large data context nowadays often refers to the number of observations n rather than to p . When it comes to applying statistical learning on such kind of big data a further buzzword often pops up — this is “deep learning”.

Deep neural networks or deep learning have attracted big attention in many fields where large data sets occur as they are promising regarding finding hidden structures and thus predicting with high accuracy; they were especially successful when features for an observation or individual can not be expressed as a single vector anymore but as a matrix, in particular that is for images [141] or features evolving and measured over time [36].

Artificial neural networks originally are inspired by neural networks in the brain [53, 117, 147]; a neural network consists of an input layer, several hidden layers, and an output layer. In the input layer the data are received. This layer is nonlinearly transformed through several hidden layers of compute units — the neurons — which are connected internally. Each of these neurons builds a weighted sum of the input of the previous layer and a nonlinear so-called activation function is applied. The last part of the neural network corresponds to the output layer which is the prediction for the new observation. As the number of hidden layers nowadays can be chosen to be large, the fashion word “deep neural network” has arisen. Deep learning is usually applied where prediction is the main goal rather than determining which variables are important [76].

Deep Neural networks have also been used in biology [3, 6]. Recently they have been used for risk prediction using Electronic Health Records (EHR) [36, 122], i.e. by using health information on patients which are electronically stored in digital format [71]. In particular, Cheng et al. [36] employs deep learning by using a matrix for each patient which contains medical events (rows) per time (columns) and incorporates this information as features in order to predict chronic diseases.

In the context of this thesis the perspectives in applying deep learning are diverse, provided that data sets with large n can be made available. First, if EHR data in the first years of life can be collected this information could be used instead of (or in addition to) cohort data and be incorporated as described above in order to predict complex diseases like the risk of childhood asthma. Second, similarly to collecting EHR data, studies on exposures can be conducted in a time-resolved fashion. For instance, exposures such as contact to

animals in the first three years of life can be recorded at regular intervals and thus again be used in a deep learning setting for the prediction of childhood asthma risk.

6.3 Conclusions

With this thesis, we contribute to methodology in disease risk prediction. We investigate the theory of the often present challenge of sample selection bias, especially by proposing a novel correction approach applicable to arbitrary machine learning algorithms and methodology for comparing performances of prediction models. In addition, we provide solutions for further challenges as complex missing data structures and high-dimensionality and we propose a strategy for prediction modeling on multi-omics data. We treat the application example of childhood asthma using modern statistical and machine learning methodology: we show how prediction for the disease can be improved and therefore identify novel predictors.

Bibliography

- [1] Animesh Acharjee, Bjorn Kloosterman, Richard G F Visser, and Chris Maliepaard. Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics*, 17(5):180, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1043-4. URL <https://doi.org/10.1186/s12859-016-1043-4>.
- [2] Aida Ali, Siti Mariyam Hj. Shamsuddin, and Anca L Ralescu. Classification with class imbalance problem: A Review. 2015.
- [3] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7):2524–2530, jul 2016. ISSN 1543-8384. doi: 10.1021/acs.molpharmaceut.6b00248. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965264/>.
- [4] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, may 2010. ISSN 1367-4803. URL <http://dx.doi.org/10.1093/bioinformatics/btq134>.
- [5] Gary P Anderson. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *The Lancet*, 372(9643):1107–1119, 2008. ISSN 0140-6736. doi: [https://doi.org/10.1016/S0140-6736\(08\)61452-X](https://doi.org/10.1016/S0140-6736(08)61452-X). URL <http://www.sciencedirect.com/science/article/pii/S014067360861452X>.
- [6] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, jul 2016. ISSN 1744-4292. doi: 10.15252/msb.20156651. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965871/>.

- [7] Donna P Ankerst, Andreas Boeck, Stephen J Freedland, J Stephen Jones, Angel M Cronin, Monique J Roobol, Jonas Hugosson, Michael W Kattan, Eric A Klein, Freddie Hamdy, David Neal, Jenny Donovan, Dipen J Parekh, Helmut Klocker, Wolfgang Horninger, Amine Benchikh, Gilles Salama, Arnauld Villers, Daniel M Moreira, Fritz H Schröder, Hans Lilja, Andrew J Vickers, and Ian M Thompson. EVALUATING THE PROSTATE CANCER PREVENTION TRIAL HIGH GRADE PROSTATE CANCER RISK CALCULATOR IN TEN INTERNATIONAL BIOSY COHORTS: RESULTS FROM THE PROSTATE BIOPSY COLLABORATIVE GROUP. *World journal of urology*, 32(1):185–191, feb 2014. ISSN 0724-4983. doi: 10.1007/s00345-012-0869-2. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3702682/>.
- [8] Donna P Ankerst, Josef Hoefler, Sebastian Bock, Phyllis J Goodman, Andrew Vickers, Javier Hernandez, Lori J Sokoll, Martin G Sanda, John T Wei, Robin J Leach, and Ian M Thompson. PROSTATE CANCER PREVENTION TRIAL RISK CALCULATOR 2.0 FOR THE PREDICTION OF LOW- VERSUS HIGH-GRADE PROSTATE CANCER. *Urology*, 83(6):1362–1368, jun 2014. ISSN 0090-4295. doi: 10.1016/j.urology.2014.02.035. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4035700/>.
- [9] M I Asher, S Montefort, B Bjorksten, C K Lai, D P Strachan, S K Weiland, H Williams, and Isaac Phase Three Study Group. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet*, 368(9537):733–743, 2006. doi: 10.1016/S0140-6736(06)69283-0. URL <http://www.ncbi.nlm.nih.gov/pubmed/16935684>[http://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(06\)69283-0/fulltext](http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(06)69283-0/fulltext).
- [10] Euan A Ashley. Towards precision medicine. *Nature Reviews Genetics*, 17:507, aug 2016. URL <http://dx.doi.org/10.1038/nrg.2016.86><http://10.0.4.14/nrg.2016.86>.
- [11] Euan A. Ashley, Basile J, VX08-770-102 Study Group, Ramsey BW, Davies J, McElvaney NG, Ross JS, Wang K, Gay L, Altman RB, Nherera L, Marks D, Minhas R, Thorogood M, Humphries SE, Blair SN, Ashley EA, Butte AJ, and Wheeler MT. The Precision Medicine Initiative. *Jama*, 313(21):2119, 2015. ISSN 0098-7484. doi: 10.1001/jama.2015.3595. URL <http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.3595>.

- [12] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple Imputation by Chained Equations: What is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, mar 2011. ISSN 1049-8931. doi: 10.1002/mpr.329. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/>.
- [13] F Backhed, J Roswall, Y Peng, Q Feng, H Jia, P Kovatcheva-Datchary, Y Li, Y Xia, H Xie, H Zhong, M T Khan, J Zhang, J Li, L Xiao, J Al-Aama, D Zhang, Y S Lee, D Kotowska, C Colding, V Tremaroli, Y Yin, S Bergman, X Xu, L Madsen, K Kristiansen, J Dahlgren, and J Wang. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe*, 17(6):852, 2015. doi: 10.1016/j.chom.2015.05.012. URL <https://www.ncbi.nlm.nih.gov/pubmed/26308884>.
- [14] Francis L Battye and Ken Shortman. Flow cytometry and cell-separation procedures. *Current Opinion in Immunology*, 3(2):239–241, 1991. ISSN 0952-7915. doi: [https://doi.org/10.1016/0952-7915\(91\)90058-9](https://doi.org/10.1016/0952-7915(91)90058-9). URL <http://www.sciencedirect.com/science/article/pii/0952791591900589>.
- [15] Daniel W Belsky, Malcolm R Sears, Robert J Hancox, Honalee Harrington, Renate Houts, Terrie E Moffitt, Karen Sugden, Benjamin Williams, Richie Poulton, and Avshalom Caspi. Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. *The Lancet. Respiratory medicine*, 1(6):453–61, aug 2013. ISSN 2213-2600. doi: 10.1016/S2213-2600(13)70101-2. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3899706&tool=pmcentrez&rendertype=abstract>.
- [16] Nicole Beydon, Stephanie D Davis, Enrico Lombardi, Julian L Allen, Hubertus G M Arets, Paul Aurora, Hans Bisgaard, G Michael Davis, Francine M Ducharme, Howard Eigen, Monika Gappa, Claude Gaultier, Per M Gustafsson, Graham L Hall, Zoltán Hantos, Michael J R Healy, Marcus H Jones, Bent Klug, Karin C Lødrup Carlsen, Sheila A McKenzie, François Marchal, Oscar H Mayer, Peter J F M Merkus, Mohy G Morris, Ellie Oostveen, J Jane Pillow, Paul C Seddon, Michael Silverman, Peter D Sly, Janet Stocks, Robert S Tepper, Daphna Vilozni, and Nicola M Wilson. An Official American Thoracic Society/European Respiratory Society Statement: Pulmonary Function Testing in Preschool Children. *American Journal of Respiratory and Critical Care Medicine*, 175(12):1304–1345, 2007. doi: 10.1164/rccm.200605-642ST. URL <http://www.atsjournals.org/doi/abs/10.1164/rccm.200605-642ST>.

- [17] C Bieli, W Eder, R Frei, C Braun-Fahrlander, W Klimecki, M Waser, J Riedler, E von Mutius, A Scheynius, G Pershagen, G Doekes, R Lauener, F D Martinez, and Parsifal study Group. A polymorphism in CD14 modifies the effect of farm milk consumption on allergic diseases and CD14 gene expression. *J Allergy Clin Immunol*, 120(6):1308–1315, 2007. doi: 10.1016/j.jaci.2007.07.034. URL <https://www.ncbi.nlm.nih.gov/pubmed/17919709>.
- [18] K Bonnelykke, P Sleiman, K Nielsen, E Kreiner-Moller, J M Mercader, D Belgrave, H T den Dekker, A Husby, A Sevelsted, G Faura-Tellez, L J Mortensen, L Paternoster, R Flaaten, A Molgaard, D E Smart, P F Thomsen, M A Rasmussen, S Bonas-Guarch, C Holst, E A Nohr, R Yadav, M E March, T Blicher, P M Lackie, V W Jaddoe, A Simpson, J W Holloway, L Duijts, A Custovic, D E Davies, D Torrents, R Gupta, M V Hollegaard, D M Hougaard, H Hakonarson, and H Bisgaard. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet*, 46(1):51–55, 2014. doi: 10.1038/ng.2830. URL <https://www.ncbi.nlm.nih.gov/pubmed/24241537>.
- [19] Leonard Bottolo and Sylvia Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.*, 5(3):583–618, 2010. ISSN 1936-0975. doi: 10.1214/10-BA523. URL <https://projecteuclid.org:443/euclid.ba/1340380542>.
- [20] Anne-Laure Boulesteix and Willi Sauerbrei. Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in bioinformatics*, 12(3):215–29, may 2011. ISSN 1477-4054. doi: 10.1093/bib/bbq085. URL <http://www.ncbi.nlm.nih.gov/pubmed/21245078>.
- [21] Anne-laure Boulesteix, Riccardo De Bin, Xiaoyu Jiang, and Mathias Fuchs. IPF-LASSO : Integrative L1 -Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine*, 2017, 2017. doi: 10.1155/2017/7691937. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5435977/pdf/CMMM2017-7691937.pdf>.
- [22] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186, 2017. ISSN 0092-8674. doi: <https://doi.org/10.1016/j.cell.2017.05.038>. URL <http://www.sciencedirect.com/science/article/pii/S0092867417306293>.
- [23] J P L Brand and J P L Brand. *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*. The Author, 1999. ISBN 9789074479080. URL <https://books.google.de/books?id=-YOTywAACAAJ>.

- [24] L Breiman, J Friedman, R Olshen, and C Stone. Classification and Regression Trees (Monterey, California: Wadsworth), 1984.
- [25] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [26] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL <http://download.springer.com/static/pdf/639/art%7D3A10.1023%7D2FA%7D3A1010933404324.pdf?originUrl=http://link.springer.com/article/10.1023/A:1010933404324%7Dtoken2=exp=1470989159%7Dac1=/static/pdf/639/art%7D253A10.1023%7D252FA%7D253A1010933404324.pdf?originUrl=http%7D3A%7D2F>.
- [27] K R Brown, R Z Krouse, A Calatroni, C M Visness, U Sivaprasad, C M Kerckmar, E C Matsui, J B West, M M Makhija, M A Gill, H Kim, M Kattan, D Pillai, J E Gern, W W Busse, A Togias, A H Liu, and G K Khurana Hershey. Endotypes of difficult-to-control asthma in inner-city African American children. *PLoS ONE*, 12(7):e0180778, jul 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0180778. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501607/>.
- [28] William S Bush and Jason H Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12):e1002822, dec 2012. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1002822. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531285/>.
- [29] William W Busse, Stephen Holgate, Edward Kerwin, Yun Chon, JingYuan Feng, Joseph Lin, and Shao-Lee Lin. Randomized, Double-Blind, Placebo-controlled Study of Brodalumab, a Human Anti-IL-17 Receptor Monoclonal Antibody, in Moderate to Severe Asthma. *American Journal of Respiratory and Critical Care Medicine*, 188(11):1294–1302, nov 2013. ISSN 1073-449X. doi: 10.1164/rccm.201212-2318OC. URL <https://doi.org/10.1164/rccm.201212-2318OC>.
- [30] Stef Van Buuren. *Flexible Imputation of Missing Data*. 2012. ISBN 9781439868249. doi: 10.1201/b11826. URL <http://bacbuc.hd.free.fr/WebDAV/data/Bouquins/vanBuuren-Flexibleimputationofmissingdata.pdf.pdf>.
- [31] Camilla Cali and Maria Longobardi. Some mathematical properties of the ROC curve and their applications. *Ricerche di Matematica*, 64(2):391–402, 2015. ISSN 1827-3491. doi: 10.1007/s11587-015-0246-8. URL <https://doi.org/10.1007/s11587-015-0246-8>.

- [32] M Caliskan, Y A Bochkov, E Kreiner-Moller, K Bonnelykke, M M Stein, G Du, H Bisgaard, D J Jackson, J E Gern, R F Lemanske Jr., D L Nicolae, and C Ober. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N Engl J Med*, 368(15):1398–1407, 2013. doi: 10.1056/NEJMoa1211592. URL <https://www.ncbi.nlm.nih.gov/pubmed/23534543>.
- [33] JOSE A. CASTRO-RODRÍGUEZ, CATHARINE J. HOLBERG, ANNE L. WRIGHT, and FERNANDO D. MARTINEZ. A Clinical Index to Define Risk of Asthma in Young Children with Recurrent Wheezing. *American Journal of Respiratory and Critical Care Medicine*, 162(4):1403–1406, oct 2000. ISSN 1073-449X. doi: 10.1164/ajrccm.162.4.9912111. URL <https://doi.org/10.1164/ajrccm.162.4.9912111>.
- [34] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [35] Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, (1999):1–12, 2004. doi: ley.edu/sites/default/files/tech-reports/666.pdf. URL <http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf>.
- [36] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. *Risk Prediction with Electronic Health Records: A Deep Learning Approach*. jun 2016. doi: 10.1137/1.9781611974348.49.
- [37] F S Collins. Shattuck lecture—medical and societal consequences of the Human Genome Project. *N Engl J Med*, 341(1):28–37, 1999. doi: 10.1056/NEJM199907013410106. URL <https://www.ncbi.nlm.nih.gov/pubmed/10387940>.
- [38] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample Selection Bias Correction Theory. page 16, 2008. doi: 10.1007/978-3-540-87987-9_8. URL <http://arxiv.org/abs/0805.2775>.
- [39] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, sep 1988. ISSN 0006-341X. doi: 10.2307/2531595. URL <http://dx.doi.org/10.2307/2531595>.
- [40] M Depner, O Fuchs, J Genuneit, A M Karvonen, A Hyvarinen, V Kaulek, C Roduit, J Weber, B Schaub, R Lauener, M Kabesch, P I Pfefferle, U Frey, J Pekkanen, J C

- Dalphin, J Riedler, C Braun-Fahrlander, E von Mutius, M J Ege, and Pasture Study Group. Clinical and epidemiologic phenotypes of childhood asthma. *Am J Respir Crit Care Med*, 189(2):129–138, 2014. doi: 10.1164/rccm.201307-1198OC. URL <https://www.ncbi.nlm.nih.gov/pubmed/24283801>.
- [41] C B Do, D A Hinds, U Francke, and N Eriksson. Comparison of family history and SNPs for predicting risk of complex disease. *PLoS Genet*, 8(10):e1002973, 2012. doi: 10.1371/journal.pgen.1002973. URL <https://www.ncbi.nlm.nih.gov/pubmed/23071447>.
- [42] F Dudbridge and A Gusnanto. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32(3):227–234, 2008. doi: 10.1002/gepi.20297. URL <https://www.ncbi.nlm.nih.gov/pubmed/18300295>.
- [43] D L Duffy, N G Martin, D Battistutta, J L Hopper, and J D Mathews. Genetics of asthma and hay fever in Australian twins. *Am Rev Respir Dis*, 142(6 Pt 1):1351–1358, 1990. doi: 10.1164/ajrccm/142.6_Pt_1.1351. URL <https://www.ncbi.nlm.nih.gov/pubmed/2252253>.
- [44] Alain Dupuy and Richard M Simon. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *JNCI: Journal of the National Cancer Institute*, 99(2):147–157, jan 2007. ISSN 0027-8874. URL <http://dx.doi.org/10.1093/jnci/djk018>.
- [45] W Eder, M J Ege, and E von Mutius. The asthma epidemic. *N Engl J Med*, 355(21):2226–2235, 2006. doi: 10.1056/NEJMra054308. URL <https://www.ncbi.nlm.nih.gov/pubmed/17124020>.
- [46] B Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 00905364. doi: 10.2307/2958830. URL <http://dx.doi.org/10.2307/2958830>.
- [47] M J Ege. Asthma and Prenatal Inflammation. *Am J Respir Crit Care Med*, 195(5):546–548, 2017. doi: 10.1164/rccm.201609-1937ED. URL <https://www.ncbi.nlm.nih.gov/pubmed/28248146>.
- [48] Markus J Ege and David P Strachan. Comparisons of power of statistical methods for gene-environment interaction analyses. *European journal of epidemiology*, 28(10):785–97, oct 2013. ISSN 1573-7284. doi: 10.1007/s10654-013-9837-4. URL <http://www.ncbi.nlm.nih.gov/pubmed/24005774>.

- [49] Markus J Ege, David P Strachan, William O C M Cookson, Miriam F Moffatt, Ivo Gut, Mark Lathrop, Michael Kabesch, Jon Genuneit, Gisela Büchele, Barbara Sozanska, Andrzej Boznanski, Paul Cullinan, Elisabeth Horak, Christian Bieli, Charlotte Braun-Fahrländer, Dick Heederik, and Erika von Mutius. Gene-environment interaction for childhood asthma and exposure to farming in Central Europe. *The Journal of allergy and clinical immunology*, 127(1):138–44, 144.e1–4, jan 2011. ISSN 1097-6825. doi: 10.1016/j.jaci.2010.09.041. URL <http://www.ncbi.nlm.nih.gov/pubmed/21211648>.
- [50] Charles Elkan. The foundations of cost-sensitive learning. *IJCAI International Joint Conference on Artificial Intelligence*, pages 973–978, 2001. ISSN 10450823. doi: doi=10.1.1.29.514. URL <http://web.cs.iastate.edu/~honavar/elkan.pdf>.
- [51] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: Models, Methods and Applications*. Springer-Verlag, Berlin, 2013.
- [52] Wei Fan and Ian Davidson. On sample selection bias and its efficient correction via model averaging and unlabeled examples. *In Proc. of SIAM Data Mining Conference*, pages 320–331, 2007. URL https://www.siam.org/proceedings/datamining/2007/dm07_{_}029fan.pdf.
- [53] B G Farley and W A Clark. Simulation of self-organizing systems by digital computer. *Trans. of the IRE Professional Group on Information Theory (TIT)*, 4:76–84, 1954.
- [54] Tom Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006. ISSN 0167-8655. doi: <http://dx.doi.org/10.1016/j.patrec.2005.10.010>. URL <http://portal.acm.org/citation.cfm?id=1159475>.
- [55] Global-Initiative for Asthma. Global strategy for the diagnosis and prevention. Global Initiative for Asthma. Last accessed May, 2017. URL <http://ginasthma.org>.
- [56] Erick Forno, Ting Wang, Qi Yan, John Brehm, Edna Acosta-Perez, Angel Colon-Semidey, Maria Alvarez, Nadia Boutaoui, Michelle M Cloutier, John F Alcorn, Glorisa Canino, Wei Chen, and Juan C Celedón. A Multiomics Approach to Identify Genes Associated with Childhood Asthma Risk and Morbidity. *American Journal of Respiratory Cell and Molecular Biology*, 57(4):439–447, 2017. doi: 10.1165/rcmb.2017-0002OC. URL <https://doi.org/10.1165/rcmb.2017-0002OC>.

- [57] Jeffrey M Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, pages 1–24, 2010. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/papers2://publication/uuid/716C673F-9976-47AB-8E0F-F42A8C7226D9>.
- [58] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 2001. ISSN 0090-5364. doi: 10.1214/aos/1013203451. URL <https://projecteuclid.org:443/euclid.aos/1013203451>.
- [59] Jerome H Friedman. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, feb 2002. ISSN 0167-9473. doi: 10.1016/S0167-9473(01)00065-2. URL [http://dx.doi.org/10.1016/S0167-9473\(01\)00065-2](http://dx.doi.org/10.1016/S0167-9473(01)00065-2).
- [60] Brigitte I. Frohnert, Michael Laimighofer, Jan Krumsiek, Fabian J. Theis, Christiane Winkler, Jill M Norris, Anette-Gabriele Ziegler, Marian J. Rewers, and Andrea K. Steck. Prediction of type 1 diabetes using a genetic risk model in the Diabetes Autoimmunity Study in the Young. *Pediatric Diabetes*, 19(2):277–283, jul 2017. ISSN 1399-543X. doi: 10.1111/pedi.12543. URL <https://doi.org/10.1111/pedi.12543>.
- [61] Mathias Fuchs and Norbert Krautenbacher. Minimization and estimation of the variance of prediction errors for cross-validation designs. *Journal of Statistical Theory and Practice*, 10(2):420–443, 2016. doi: 10.1080/15598608.2016.1158675. URL <http://dx.doi.org/10.1080/15598608.2016.1158675>.
- [62] Mathias Fuchs, Tim Beißbarth, Edgar Wingender, and Klaus Jung. Connecting high-dimensional mRNA and miRNA expression data for binary medical classification problems. *Computer Methods and Programs in Biomedicine*, 111(3):592–601, 2013. ISSN 0169-2607. doi: <https://doi.org/10.1016/j.cmpb.2013.05.013>. URL <http://www.sciencedirect.com/science/article/pii/S0169260713001703>.
- [63] Oliver Fuchs, Jon Genuneit, Philipp Latzin, Gisela Büchele, Elisabeth Horak, Georg Loss, Barbara Sozanska, Juliane Weber, Andrzej Boznanski, Dick Heederik, Charlotte Braun-Fahrlander, Urs Frey, and Erika Von Mutius. Farming environments and childhood atopy, wheeze, lung function, and exhaled nitric oxide. *Journal of Allergy and Clinical Immunology*, 130(2), 2012. ISSN 00916749. doi: 10.1016/j.jaci.2012.04.049. URL http://ac.els-cdn.com/S0091674912007889/1-s2.0-S0091674912007889-main.pdf?{}_tid=a3e9d238-d149-11e6-8c5e-00000aab0f6b{&}acdnat=1483402669{ }f3fceb3d85f967d6497eb6445d7ce86d.

- [64] Consortium Genomes Project, G R Abecasis, A Auton, L D Brooks, M A DePristo, R M Durbin, R E Handsaker, H M Kang, G T Marth, and G A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422): 56–65, 2012. doi: 10.1038/nature11632. URL <https://www.ncbi.nlm.nih.gov/pubmed/23128226>.
- [65] J Genuneit, G Buchele, M Waser, K Kovacs, A Debinska, A Boznanski, C Strunz-Lehner, E Horak, P Cullinan, D Heederik, C Braun-Fahrlander, E von Mutius, and Gabriela Study Group. The GABRIEL Advanced Surveys: study design, participation and evaluation of bias. *Paediatr Perinat Epidemiol*, 25(5):436–447, 2011. doi: 10.1111/j.1365-3016.2011.01223.x. URL <https://www.ncbi.nlm.nih.gov/pubmed/21819425>.
- [66] Jon Genuneit. To the Editor: Sex-Specific Development of Asthma Differs between Farm and Nonfarm Children: A Cohort Study. *American Journal of Respiratory and Critical Care Medicine*, 190(5):588–590, 2014.
- [67] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.531. URL <http://science.sciencemag.org/content/286/5439/531>.
- [68] S G Gregory, K F Barlow, K E McLay, R Kaul, D Swarbreck, A Dunham, C E Scott, K L Howe, K Woodfine, C C A Spencer, M C Jones, C Gillson, S Searle, Y Zhou, F Kokocinski, L McDonald, R Evans, K Phillips, A Atkinson, R Cooper, C Jones, R E Hall, T D Andrews, C Lloyd, R Ainscough, J P Almeida, K D Ambrose, F Anderson, R W Andrew, R I S Ashwell, K Aubin, A K Babbage, C L Bagguley, J Bailey, H Beasley, G Bethel, C P Bird, S Bray-Allen, J Y Brown, A J Brown, D Buckley, J Burton, J Bye, C Carder, J C Chapman, S Y Clark, G Clarke, C Clee, V Copley, R E Collier, N Corby, G J Coville, J Davies, R Deadman, M Dunn, M Earthrowl, A G Ellington, H Errington, A Frankish, J Frankland, L French, P Garner, J Garnett, L Gay, M R J Ghori, R Gibson, L M Gilby, W Gillett, R J Glithero, D V Grafham, C Griffiths, S Griffiths-Jones, R Grocock, S Hammond, E S I Harrison, E Hart, E Haugen, P D Heath, S Holmes, K Holt, P J Howden, A R Hunt, S E Hunt, G Hunter, J Isherwood, R James, C Johnson, D Johnson, A Joy, M Kay, J K Kershaw, M Kibukawa, A M Kimberley, A King, A J Knights, H Lad, G Laird, S Lawlor, D A Leongamornlert, D M Lloyd, J Loveland, J Lovell, M J

- Lush, R Lyne, S Martin, M Mashreghi-Mohammadi, L Matthews, N S W Matthews, S McLaren, S Milne, S Mistry, M J F Moore, T Nickerson, C N O'Dell, K Oliver, A Palmeiri, S A Palmer, A Parker, D Patel, A V Pearce, A I Peck, S Pelan, K Phelps, B J Phillimore, R Plumb, J Rajan, C Raymond, G Rouse, C Saenphimmachak, H K Sehra, E Sheridan, R Shownkeen, S Sims, C D Skuce, M Smith, C Steward, S Subramanian, N Sycamore, A Tracey, A Tromans, Z Van Helmond, M Wall, J M Wallis, S White, S L Whitehead, J E Wilkinson, D L Willey, H Williams, L Wilming, P W Wray, Z Wu, A Coulson, M Vaudin, J E Sulston, R Durbin, T Hubbard, R Wooster, I Dunham, N P Carter, G McVean, M T Ross, J Harrow, M V Olson, S Beck, J Rogers, and D R Bentley. The DNA sequence and biological annotation of human chromosome 1. *Nature*, 441:315, may 2006. URL <http://dx.doi.org/10.1038/nature04727><http://10.0.4.14/nature04727><https://www.nature.com/articles/nature04727#supplementary-information>.
- [69] D A Grimes and K F Schulz. Refining clinical diagnosis with likelihood ratios. *Lancet*, 365(9469):1500–1505, 2005. doi: 10.1016/S0140-6736(05)66422-7. URL <https://www.ncbi.nlm.nih.gov/pubmed/15850636>.
- [70] Justin Guinney, Tao Wang, Teemu D Laajala, Kimberly Kanigel Winner, J Christopher Bare, Elias Chaibub Neto, Suleiman A Khan, Gopal Peddinti, Antti Airola, Tapio Pahikkala, Tuomas Mirtti, Thomas Yu, Brian M Bot, Liji Shen, Kald Abdallah, Thea Norman, Stephen Friend, Gustavo Stolovitzky, Howard Soule, Christopher J Sweeney, Charles J Ryan, Howard I Scher, Oliver Sartor, Yang Xie, Tero Aittokallio, Fang Liz Zhou, James C Costello, Kald Abdallah, Tero Aittokallio, Antti Airola, Catalina Anghe, Helia Azima, Robert Baertsch, Pedro J Ballester, Chris Bare, Vinayak Bhandari, Brian M Bot, Cuong C Dang, Maria Bekker-Nielsen Dunbar, Ann-Sophie Buchardt, Ljubomir Buturovic, Da Cao, Prabhakar Chalise, Junwoo Cho, Tzu-Ming Chu, R Yates Coley, Sailesh Conjeti, Sara Correia, James C Costello, Ziwei Dai, Junqiang Dai, Philip Dargatz, Sam Delavarkhan, Detian Deng, Ankur Dhanik, Yu Du, Aparna Elangovan, Shellie Ellis, Laura L Elo, Shadrielle M Espiritu, Fan Fan, Ashkan B Farshi, Ana Freitas, Brooke Fridley, Stephen Friend, Christiane Fuchs, Eyal Gofer, Gopalacharyulu Peddinti, Stefan Graw, Russ Greiner, Yuanfang Guan, Justin Guinney, Jing Guo, Pankaj Gupta, Anna I Guyer, Jiawei Han, Niels R Hansen, Billy H W Chang, Outi Hirvonen, Barbara Huang, Chao Huang, Jinseub Hwang, Joseph G Ibrahim, Vivek Jayaswa, Jouhyun Jeon, Zhicheng Ji, Deekshith Juvvadi, Sirkku Jyrkkiö, Kimberly Kanigel-Winner, Amin Katouzian, Marat D Kazanov, Suleiman A Khan, Shahin Khayyer, Dalho Kim, Agnieszka K Golinska, Devin Koestler, Fernanda Kokowicz, Ivan Kondofersky, Norbert Krauten-

- bacher, Damjan Krstajic, Luke Kumar, Christoph Kurz, Matthew Kyan, Teemu D Laajala, Michael Laimighofer, Eunjee Lee, Wojciech Lesinski, Miao Zhu Li, Ye Li, Qiuyu Lian, Xiaotao Liang, Minseong Lim, Henry Lin, Xihui Lin, Jing Lu, Mehrad Mahmoudian, Roozbeh Manshaei, Richard Meier, Dejan Miljkovic, Tuomas Mirtti, Krzysztof Mnich, Nassir Navab, Elias C Neto, Yulia Newton, Thea Norman, Tapio Pahikkala, Subhabrata Pal, Byeongju Park, Jaykumar Patel, Swetabh Pathak, Alejandrina Pattin, Donna P Ankerst, Jian Peng, Anne H Petersen, Robin Philip, Stephen R Piccolo, Sebastian Pölsterl, Aneta Polewko-Klim, Karthik Rao, Xiang Ren, Miguel Rocha, Witold R Rudnicki, Charles J Ryan, Hyunnam Ryu, Oliver Sartor, Hagen Scherb, Raghav Sehgal, Fatemeh Seyednasrollah, Jingbo Shang, Bin Shao, Liji Shen, Howard Sher, Motoki Shiga, Artem Sokolov, Julia F Söllner, Lei Song, Howard Soule, Gustavo Stolovitzky, Josh Stuart, Ren Sun, Christopher J Sweeney, Nazanin Tahmasebi, Kar-Tong Tan, Lisbeth Tomaziu, Joseph Usset, Yee-leng S Vang, Roberto Vega, Vitor Vieira, David Wang, Difei Wang, Junmei Wang, Lichao Wang, Sheng Wang, Tao Wang, Yue Wang, Russ Wolfinger, Chris Wong, Zhenke Wu, Jinfeng Xiao, Xiaohui Xie, Yang Xie, Doris Xin, Hojin Yang, Nancy Yu, Thomas Yu, Xiang Yu, Sulmaz Zahedi, Massimiliano Zanin, Chihao Zhang, Jingwen Zhang, Shihua Zhang, Yanchun Zhang, Fang Liz Zhou, Hongtu Zhu, Shan-feng Zhu, and Yuxin Zhu. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. *The Lancet Oncology*, 18(1): 132–142, feb 2018. ISSN 1470-2045. doi: 10.1016/S1470-2045(16)30560-5. URL [http://dx.doi.org/10.1016/S1470-2045\(16\)30560-5](http://dx.doi.org/10.1016/S1470-2045(16)30560-5).
- [71] Tracy D Gunter and Nicolas P Terry. The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. *Journal of Medical Internet Research*, 7(1):e3, mar 2005. ISSN 1438-8871. doi: 10.2196/jmir.7.1.e3. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1550638/>.
- [72] V P Guntur and C R Reiner. The potential use of tyrosine kinase inhibitors in severe asthma. *Curr Opin Allergy Clin Immunol*, 12(1):68–75, 2012. doi: 10.1097/ACI.0b013e32834ecb4f. URL <http://www.ncbi.nlm.nih.gov/pubmed/22157153>.
- [73] Alan E Guttmacher, Amy L McGuire, Bruce Ponder, and Kári Stefánsson. Personalized genomic information: preparing for the future of genetic medicine. *Nature Reviews Genetics*, 11:161, jan 2010. URL <http://dx.doi.org/10.1038/nrg2735><http://10.0.4.14/nrg2735>.

- [74] I Guyon. A scaling law for the validation-set training-set size ratio, 1997. URL [citeulike-article-id:2711605http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1337](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1337).
- [75] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, apr 1982. ISSN 0033-8419. doi: 10.1148/radiology.143.1.7063747. URL <https://doi.org/10.1148/radiology.143.1.7063747>.
- [76] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- [77] James J Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979. URL <http://faculty.smu.edu/millimet/classes/eco7321/papers/heckman02.pdf>.
- [78] Eva Hedlund and Qiaolin Deng. Single-cell RNA sequencing: Technical advancements and biological applications. *Molecular Aspects of Medicine*, 59:36–46, 2018. ISSN 0098-2997. doi: <https://doi.org/10.1016/j.mam.2017.07.003>. URL <http://www.sciencedirect.com/science/article/pii/S0098299717300535>.
- [79] Stefanie Hieke, Axel Benner, Richard F Schlenk, Martin Schumacher, and Harald Binder. Integrating multiple molecular sources into a clinical risk prediction signature by extracting complementary information. *BMC Bioinformatics*, pages 1–24, 2016. ISSN 14712105. doi: 10.1186/s12859-016-1183-6. URL <http://dx.doi.org/10.1186/s12859-016-1183-6http://download.springer.com/static/pdf/204/art%}3A10.1186%}2Fs12859-016-1183-6.pdf?originUrl=http://bmcbioinformatics.biomedcentral.com/article/10.1186/s12859-016-1183-6%}token2=exp=1475238511%}ac1=/static/pdf/204/>.
- [80] E Horak, B Morass, H Ulmer, J Genuneit, C Braun-Fahrlander, E von Mutius, and Gabriel Study Group. Prevalence of wheezing and atopic diseases in Austrian schoolchildren in conjunction with urban, rural or farm residence. *Wien Klin Wochenschr*, 126(17-18):532–536, 2014. doi: 10.1007/s00508-014-0571-z. URL <https://www.ncbi.nlm.nih.gov/pubmed/25047409>.
- [81] Nicholas J Horton and Ken P Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American statistician*, 61(1):79–90, feb 2007. ISSN 0003-1305. doi:

- 10.1198/000313007X172556. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839993/>.
- [82] D.G. Horvitz and D.J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. URL <http://lib.stat.cmu.edu/~brian/905-2008/papers/Horvitz-Thompson-1952-jasa.pdf>.
- [83] Jiayuan Huang, Alexander J Smola, Arthur Gretton, Karsten M Borgwardt, and Bernhard Scholkopf. Correcting Sample Selection Bias by Unlabeled Data. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 601–608, Cambridge, MA, USA, 2006. MIT Press. URL <http://dl.acm.org/citation.cfm?id=2976456.2976532>.
- [84] Ying Huang and Margaret Sullivan Pepe. Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods. *Statistics in medicine*, 29(13):1391–410, 2010. ISSN 1097-0258. doi: 10.1002/sim.3876. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3045657&tool=pmcentrez&rendertype=abstract>.
- [85] Huber, W., Carey, V J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B S., Bravo, H C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K D., Irizarry, R A., Lawrence, M., Love, M I., MacDonald, J., Obenchain, V., Ole's, A K., Pag'es, H., Reyes, A., Shannon, P., Smyth, G K., Tenenbaum, D., Waldron, L., Morgan, and M. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015. URL <http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html>.
- [86] Michelle Fox Huffaker and Wanda Phipatanakul. Utility of the Asthma Predictive Index in predicting childhood asthma and identifying disease-modifying interventions. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology*, 112(3):188–190, mar 2014. ISSN 1081-1206. doi: 10.1016/j.anai.2013.12.001. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4019987/>.
- [87] Zou Hui and Hastie Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, mar 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.

- [88] S Illi, M Depner, J Genuneit, E Horak, G Loss, C Strunz-Lehner, G Buchele, A Boznanski, H Danielewicz, P Cullinan, D Heederik, C Braun-Fahrlander, E von Mutius, and Gabriela Study Group. Protection from childhood asthma and allergy in Alpine farm environments-the GABRIEL Advanced Studies. *J Allergy Clin Immunol*, 129(6):1470–7 e6, 2012. doi: 10.1016/j.jaci.2012.03.013. URL <https://www.ncbi.nlm.nih.gov/pubmed/22534534>.
- [89] Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data, nov 2016. URL <https://doi.org/10.1007/s11634-016-0276-4>.
- [90] Kristel J M Janssen, Yvonne Vergouwe, Cor J Kalkman, Diederick E Grobbee, and Karel G M Moons. A simple method to adjust clinical prediction models to local circumstances. *Canadian journal of anaesthesia = Journal canadien d’anesthésie*, 56(3):194–201, 2009. ISSN 0832-610X. doi: 10.1007/s12630-009-9041-x. URL <http://www.ncbi.nlm.nih.gov/pubmed/19247740>.
- [91] Nathalie Japkowicz. The Class Imbalance Problem: Significance and Strategies. In *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, pages 111–117, 2000.
- [92] Monika Jelizarow, Vincent Guillemot, Arthur Tenenhaus, Korbinian Strimmer, and Anne-Laure Boulesteix. Over-optimism in bioinformatics: an illustration. *Bioinformatics*, 26(16):1990–1998, aug 2010. ISSN 1367-4803. URL <http://dx.doi.org/10.1093/bioinformatics/btq323>.
- [93] Jia Kang, Judy Cho, and Hongyu Zhao. PRACTICAL ISSUES IN BUILDING RISK-PREDICTING MODELS FOR COMPLEX DISEASES. *Journal of biopharmaceutical statistics*, 20(2):415–440, mar 2010. ISSN 1054-3406. doi: 10.1080/10543400903572829. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3685822/>.
- [94] Robert E Kass and Adrian E Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572. URL <http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572>.
- [95] Carl Kendall, Ligia R F S Kerr, Rogerio C Gondim, Guilherme L Werneck, Raimunda Hermelinda Maia Macena, Marta Kerr Pontes, Lisa G Johnston, Keith Sabin, and Willi McFarland. An Empirical Comparison of Respondent-driven Sampling, Time Location Sampling, and Snowball Sampling for Behavioral Surveil-

- lance in Men Who Have Sex with Men, Fortaleza, Brazil. *AIDS and Behavior*, 12(1):97, 2008. ISSN 1573-3254. doi: 10.1007/s10461-008-9390-4. URL <http://dx.doi.org/10.1007/s10461-008-9390-4>.
- [96] C D Kidd, P J Thompson, L Barrett, and S Baltic. Histone Modifications and Asthma. The Interface of the Epigenetic and Genetic Landscapes. *Am J Respir Cell Mol Biol*, 54(1):3–12, 2016. doi: 10.1165/rcmb.2015-0050TR. URL <https://www.ncbi.nlm.nih.gov/pubmed/26397168>.
- [97] Gary King and Langche Zeng. Logistic Regression in Rare Events Data. *Political Analysis*, 9:137–163, 2001.
- [98] Ivan Kondofersky, Michael Laimighofer, Christoph Kurz, Norbert Krautenbacher, Julia Söllner, Philip Dargatz, Donna Ankerst, and Christiane Fuchs. Three general concepts to improve risk prediction: good data, wisdom of the crowd, recalibration. *F1000Research*, 5:2671, 2016. doi: 10.12688/f1000research.8680.1.
- [99] Maritha J Kotze, Hilmar K Lückhoff, Armand V Peeters, Karin Baatjes, Mardelle Schoeman, Lize van der Merwe, Kathleen A Grant, Leslie R Fisher, Nicole van der Merwe, Jacobus Pretorius, David P van Velden, Ettienne J Myburgh, Fredrieka M Pienaar, Susan J van Rensburg, Yandiswa Y Yako, Alison V September, Kelebogile E Moremi, Frans J Cronje, Nicki Tiffin, Christianne S H Bouwens, Juanita Bezuidenhout, Justus P Appfelstaedt, F Stephen Hough, Rajiv T Erasmus, and Johann W Schneider. Genomic medicine and risk prediction across the disease spectrum. *Critical Reviews in Clinical Laboratory Sciences*, 52(3):120–137, may 2015. ISSN 1040-8363. doi: 10.3109/10408363.2014.997930. URL <https://doi.org/10.3109/10408363.2014.997930>.
- [100] Norbert Krautenbacher, Fabian J. Theis, and Christiane Fuchs. Correcting Classifiers for Sample Selection Bias in Two-Phase Case-Control Studies. *Computational and Mathematical Methods in Medicine*, 2017:18, 2017. doi: 10.1155/2017/7847531.
- [101] Norbert Krautenbacher, Nicolai Flach, Andreas Böck, Kristina Laubhahn, Michael Laimighofer, Fabian J Theis, Donna P Ankerst, Christiane Fuchs, and Bianca Schaub. Classifying childhood asthma phenotypes from genetic, immunological and environmental factors: A strategy for high-dimensional multivariable analysis. *submitted*, 2018.
- [102] Norbert Krautenbacher, Michael Kabesch, Elisabeth Horak, Charlotte Braun-Fahländer, Jon Genuit, Andrzej Boznanski, Erika von Mutius, Fabian J Theis,

- Christiane Fuchs, and Markus J Ege. Predicting childhood asthma risk by genetic and environmental variables. *submitted*, 2018.
- [103] Norbert Krautenbacher, Kevin Strauss, Maximilian Mandl, and Christiane Fuchs. *sambia: A Collection of Techniques Correcting for Sample Selection Bias*, 2018. URL <https://cran.r-project.org/package=sambia>.
- [104] Jochen Kruppa, Andreas Ziegler, and Inke R. König. Risk estimation and risk prediction using machine-learning methods. *Human Genetics*, 131(10):1639–1654, 2012. ISSN 03406717. doi: 10.1007/s00439-012-1194-y. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3432206/pdf/439{}_2012{}_Article{}_1194.pdf.
- [105] Miroslav Kubat and Stan Matwin. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.
- [106] Michael Laimighofer, Jan Krumsiek, Florian Buettner, and Fabian J. Theis. Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression. *Journal of Computational Biology*, 23(4):cmb.2015.0192, 2016. ISSN 1066-5277. doi: 10.1089/cmb.2015.0192. URL <http://online.liebertpub.com/doi/10.1089/cmb.2015.0192>.
- [107] Katja Landgraf-Rauf, Bettina Anselm, and Bianca Schaub. The puzzle of immune phenotypes of childhood asthma. *Molecular and Cellular Pediatrics*, 3(1):27, 2016. ISSN 2194-7791. doi: 10.1186/s40348-016-0057-3. URL <https://doi.org/10.1186/s40348-016-0057-3>.
- [108] David D Lewis and Jason Catlett. Heterogeneous Uncertainty Sampling for Supervised Learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1994.
- [109] X Li, T D Howard, S L Zheng, T Haselkorn, S P Peters, D A Meyers, and E R Bleeker. Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. *J Allergy Clin Immunol*, 125(2):328–335 e11, 2010. doi: 10.1016/j.jaci.2009.11.018. URL <https://www.ncbi.nlm.nih.gov/pubmed/20159242>.
- [110] Y Li, C J Willer, J Ding, P Scheet, and G R Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–834, 2010. doi: 10.1002/gepi.20533. URL <https://www.ncbi.nlm.nih.gov/pubmed/21058334>.

- [111] Tang Lingqi, Song Juwon, Belin Thomas R., and Unützer Jürgen. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24(14):2111–2128, may 2005. ISSN 0277-6715. doi: 10.1002/sim.2099. URL <https://doi.org/10.1002/sim.2099>.
- [112] Roderick J A Little. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988. ISSN 07350015. doi: 10.2307/1391878. URL <http://www.jstor.org/stable/1391878>.
- [113] Thomas Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(8):1–19, 2004. ISSN 15487660. URL <http://www.doaj.org/doaj?func=fulltext&aId=88360>.
- [114] J MacArthur, E Bowler, M Cerezo, L Gil, P Hall, E Hastings, H Junkins, A McMahon, A Milano, J Morales, Z M Pendlington, D Welter, T Burdett, L Hindorff, P Flicek, F Cunningham, and H Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*, 45(D1):D896–D901, 2017. doi: 10.1093/nar/gkw1133. URL <https://www.ncbi.nlm.nih.gov/pubmed/27899670>.
- [115] I Marenholz, J Esparza-Gordillo, F Ruschendorf, A Bauerfeind, D P Strachan, B D Spycher, H Baurecht, P Margaritte-Jeannin, A Saaf, M Kerkhof, M Ege, S Baltic, M C Matheson, J Li, S Michel, W Q Ang, W McArdle, A Arnold, G Homuth, F Demenais, E Bouzigon, C Soderhall, G Pershagen, J C de Jongste, D S Postma, C Braun-Fahrlander, E Horak, L M Ogorodova, V P Puzyrev, E Y Bragina, T J Hudson, C Morin, D L Duffy, G B Marks, C F Robertson, G W Montgomery, B Musk, P J Thompson, N G Martin, A James, P Sleiman, E Toskala, E Rodriguez, R Folster-Holst, A Franke, W Lieb, C Gieger, A Heinzmann, E Rietschel, T Keil, S Cichon, M M Nothen, C E Pennell, P D Sly, C O Schmidt, A Matanovic, V Schneider, M Heinig, N Hubner, P G Holt, S Lau, M Kabesch, S Weidinger, H Hakonarson, M A Ferreira, C Laprise, M B Freidin, J Genuneit, G H Koppelman, E Melen, M H Dizier, A J Henderson, and Y A Lee. Meta-analysis identifies seven susceptibility loci involved in the atopic march. *Nat Commun*, 6:8804, 2015. doi: 10.1038/ncomms9804. URL <https://www.ncbi.nlm.nih.gov/pubmed/26542096>.
- [116] Kimberly McAllister, Leah E Mechanic, Christopher Amos, Hugues Aschard, Ian A Blair, Nilanjan Chatterjee, David Conti, W James Gauderman, Li Hsu, Carolyn M Hutter, Marta M Jankowska, Jacqueline Kerr, Peter Kraft, Stephen B Montgomery, Bhramar Mukherjee, George J Papanicolaou, Chirag J Patel, Marylyn D Ritchie, Beate R Ritz, Duncan C Thomas, Peng Wei, John S Witte,

- and on behalf of workshop Participants. Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *American Journal of Epidemiology*, 186(7):753–761, oct 2017. ISSN 0002-9262. URL <http://dx.doi.org/10.1093/aje/kwx227>.
- [117] Warren S McCulloch and Walter Pitts. Neurocomputing: Foundations of Research. chapter A Logical, pages 15–27. MIT Press, Cambridge, MA, USA, 1988. ISBN 0-262-01097-6. URL <http://dl.acm.org/citation.cfm?id=65669.104377>.
- [118] Sterling McPherson, Celestina Barbosa-Leiker, Mary Rose Mamey, Michael McDonnell, Craig K Enders, and John Roll. A ‘Missing Not at Random’ (MNAR) and ‘Missing at Random’ (MAR) Growth Model Comparison with a Buprenorphine/Naloxone Clinical Trial. *Addiction (Abingdon, England)*, 110(1):51–58, jan 2015. ISSN 0965-2140. doi: 10.1111/add.12714. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270922/>.
- [119] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*, TU Wien, 2015. URL <http://cran.r-project.org/package=e1071>.
- [120] S Michel, F Busato, J Genuneit, J Pekkanen, J C Dalphin, J Riedler, N Mazaleyrat, J Weber, A M Karvonen, M R Hirvonen, C Braun-Fahrlander, R Lauener, E von Mutius, M Kabesch, J Tost, and Pasture study Group. Farm exposure and time trends in early childhood may influence DNA methylation in genes related to asthma and allergy. *Allergy*, 68(3):355–364, 2013. doi: 10.1111/all.12097. URL <https://www.ncbi.nlm.nih.gov/pubmed/23346934>.
- [121] T. C. Mills, R. Stall, L. Pollack, J. P. Paul, D. Binson, J. Canchola, and J. A. Catania. Health-related characteristics of men who have sex with men: A comparison of those living in "gay ghettos" with those living elsewhere. *American Journal of Public Health*, 91(6):980–983, 2001. ISSN 00900036. doi: 10.2105/AJPH.91.6.980.
- [122] Riccardo Miotto, li Li, and Joel T. Dudley. *Deep Learning to Predict Patient Future Diseases from the Electronic Health Records*, volume 9626. mar 2016. ISBN 978-3-319-30670-4. doi: 10.1007/978-3-319-30671-1_66.
- [123] Miriam F Moffatt, Ivo G Gut, Florence Demenais, David P Strachan, Emmanuelle Bouzigon, Simon Heath, Erika von Mutius, Martin Farrall, Mark Lathrop, and William O C M Cookson. A large-scale, consortium-based genomewide association study of asthma. *The New England journal of medicine*, 363(13):1211–

- 21, sep 2010. ISSN 1533-4406. doi: 10.1056/NEJMoa0906312. URL <http://www.ncbi.nlm.nih.gov/pubmed/20860503>.
- [124] M Mommers, G M H Swaen, M Weishoff-Houben, W Dott, and C P van Schayck. Differences in asthma diagnosis and medication use in children living in Germany and the Netherlands. *Primary Care Respiratory Journal*, 14(1):31–37, 2005.
- [125] Daniel J Mundfrom, Adam Piccone, Jamis J Perrett, Jay Schaffer, and Michelle Roozeboom. Bonferroni Adjustments in Tests for Regression Coefficients. *Multiple Linear Regression Viewpoints*, 32:1–6, 2006. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.490.7640{&}rep=rep1{&}type=pdf>.
- [126] Matthew Nahorniak, David P Larsen, Carol Volk, and Chris E Jordan. Using Inverse Probability Bootstrap Sampling to Eliminate Sample Induced Bias in Model Based Analysis of Unequal Probability Samples. *PLOS ONE*, 10(6):1–19, 2015. doi: 10.1371/journal.pone.0131765. URL <http://dx.doi.org/10.1371{&}2Fjournal.pone.0131765>.
- [127] Parameswaran Nair, Sally Wenzel, Klaus F Rabe, Arnaud Bourdin, Njira L Lugogo, Piotr Kuna, Peter Barker, Stephanie Sproule, Sandhia Ponnarambil, and Mitchell Goldman. Oral Glucocorticoid-Sparing Effect of Benralizumab in Severe Asthma. *New England Journal of Medicine*, 376(25):2448–2458, may 2017. ISSN 0028-4793. doi: 10.1056/NEJMoa1703501. URL <https://doi.org/10.1056/NEJMoa1703501>.
- [128] Paul J Newcombe, David V Conti, and Sylvia Richardson. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genetic Epidemiology*, 40(3):188–201, mar 2016. ISSN 0741-0395. doi: 10.1002/gepi.21953. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817278/>.
- [129] C Ober. Asthma Genetics in the Post-GWAS Era. *Ann Am Thorac Soc*, 13 Suppl 1: S85–90, 2016. doi: 10.1513/AnnalsATS.201507-459MG. URL <https://www.ncbi.nlm.nih.gov/pubmed/27027959>.
- [130] C Ober and S Hoffjan. Asthma genetics 2006: the long and winding road to gene discovery. *Genes and immunity*, 7(2):95–100, mar 2006. ISSN 1466-4879. doi: 10.1038/sj.gene.6364284. URL <http://www.ncbi.nlm.nih.gov/pubmed/16395390>.
- [131] C Ober and D Vercelli. Gene-environment interactions in human disease: nuisance or opportunity? *Trends Genet*, 27(3):107–115, 2011. doi: 10.1016/j.tig.2010.12.004. URL <https://www.ncbi.nlm.nih.gov/pubmed/21216485>.

- [132] Joseph O Ogutu, Hans-Peter Piepho, and Torben Schulz-Streeck. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, 5(Suppl 3):S11–S11, may 2011. ISSN 1753-6561. doi: 10.1186/1753-6561-5-S3-S11. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103196/>.
- [133] M Pak and M Shin. Developing disease risk prediction model based on environmental factors. In *The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014)*, pages 1–2, 2014. ISBN 0747-668X VO -. doi: 10.1109/ISCE.2014.6884338.
- [134] J T Parsons and S J Parsons. Src family protein tyrosine kinases: cooperating with growth factor and adhesion signaling pathways. *Curr Opin Cell Biol*, 9(2):187–192, 1997. URL <http://www.ncbi.nlm.nih.gov/pubmed/9069259>.
- [135] B Pasaniuc and A L Price. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*, 18(2):117–127, 2017. doi: 10.1038/nrg.2016.142. URL <https://www.ncbi.nlm.nih.gov/pubmed/27840428>.
- [136] Corrado Pelaia, Alessandro Vatrella, Andrea Bruni, Rosa Terracciano, and Girolamo Pelaia. Benralizumab in the treatment of severe asthma: design, development and potential place in therapy. *Drug Design, Development and Therapy*, 12:619–628, mar 2018. ISSN 1177-8881. doi: 10.2147/DDDT.S155307. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5868576/>.
- [137] Brand Jaap P.L., Buuren Stef, Groothuis-Oudshoorn Karin, and Gelsema Edzard S. A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica*, 57(1):36–45, may 2003. ISSN 0039-0402. doi: 10.1111/1467-9574.00219. URL <https://doi.org/10.1111/1467-9574.00219>.
- [138] Foster Provost and Pedro Domingos. Well-trained PETs: Improving probability estimation trees. *CeDER Working Paper #IS-00-04*, Stern School of Business, New York University, NY, NY 10012, 2001.
- [139] Diana Raedler, Nikolaus Ballenberger, Elisabeth Klucker, Andreas Böck, Ragna Otto, Olivia Prazeres Da Costa, Otto Holst, Thomas Illig, Thorsten Buch, Erika Von Mutius, and Bianca Schaub. Identification of novel immune phenotypes for allergic and nonallergic childhood asthma. *Journal of Allergy and Clinical Immunology*, 135(1):81–91, 2015. ISSN 10976825. doi: 10.1016/j.jaci.2014.07.046.
- [140] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27, 2001.

- [141] Waseem Rawat and Zenghui Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural computation*, 29 9:2352–2449, 2017.
- [142] Greg Ridgeway. Generalized Boosted Models : A guide to the gbm package. (4): 1–12, 2007. URL <http://www.saedsayad.com/docs/gbm2.pdf>.
- [143] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12(1):77, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-77. URL <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-77>.
- [144] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 01621459. URL <http://www.jstor.org/stable/2290910>.
- [145] S Romanet-Manent, D Charpin, A Magnan, A Lanteaume, D Vervloet, and Egea Cooperative Group. Allergic vs nonallergic asthma: what makes the difference? *Allergy*, 57(7):607–613, 2002. URL <http://www.ncbi.nlm.nih.gov/pubmed/12100301>.
- [146] Sherri Rose and MJ van der Laan. A Note on Risk Prediction for Case-Control Studies. [*Preprint*], 2008. URL <http://biostats.bepress.com/ucbbiostat/paper241/>.
- [147] Frank Rosenblatt. The Perceptron: {A} Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65:386–408, 1958.
- [148] D B Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.
- [149] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581. URL [+http://dx.doi.org/10.1093/biomet/63.3.581](http://dx.doi.org/10.1093/biomet/63.3.581).
- [150] Olli Saarela, Sangita Kulathinal, and Juha Karvanen. Secondary Analysis under Cohort Sampling Designs Using Conditional Likelihood. *Journal of Probability and Statistics*, 2012(35):1–37, 2012.
- [151] Tobi Saidel, Rajatashuvra Adhikary, Mandar Mainkar, Jayesh Dale, Virginia Loo, Motiur Rahman, Banadakoppa M Ramesh, and Ramesh S Paranjape. Baseline integrated behavioural and biological assessment among most at-risk populations in six high-prevalence states of India: design and implementation challenges. *AIDS (London, England)*, 22 Suppl 5(November):S17–S34, 2008. ISSN 0269-9370. doi: 10.1097/01.aids.0000343761.77702.04.

- [152] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. URL <http://lmdvr.r-forge.r-project.org>.
- [153] Jaya M Satagopan, E S Venkatraman, and Colin B Begg. Two-Stage Designs for Gene – Disease Association Studies with Sample Size Constraints. *Biometrics*, 60 (September):589–597, 2004.
- [154] James J Schlesselman and Paul D Stolley. *Case-control studies : design, conduct, analysis*. Oxford University Press, Oxford, UK, 1982.
- [155] P C Schroder, V I Casaca, S Illi, M Schieck, S Michel, A Bock, C Roduit, R Frei, A Lluís, J Genuneit, P Pfefferle, M Roponen, J Weber, C Braun-Fahrlander, J Riedler, R Lauener, D A Vuitton, J C Dalphin, J Pekkanen, E von Mutius, M Kabesch, B Schaub, and Pasture Study Group. IL-33 polymorphisms are associated with increased risk of hay fever and reduced regulatory T cells in a birth cohort. *Pediatr Allergy Immunol*, 27(7):687–695, 2016. doi: 10.1111/pai.12597. URL <https://www.ncbi.nlm.nih.gov/pubmed/27171815>.
- [156] Johanna M Seddon, Robyn Reynolds, Julian Maller, Jesen A Fagerness, Mark J Daly, and Bernard Rosner. Prediction Model for Prevalence and Incidence of Advanced Age-Related Macular Degeneration Based on Genetic, Demographic, and Environmental Variables. *Investigative ophthalmology & visual science*, 50 (5):2044–2053, may 2009. ISSN 0146-0404. doi: 10.1167/iovs.08-3064. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3772781/>.
- [157] R Sehmi, S G Smith, M Kjarsgaard, K Radford, L P Boulet, C Lemiere, C M Prazma, H Ortega, J G Martin, and P Nair. Role of local eosinophilopoietic processes in the development of airway eosinophilia in prednisone-dependent severe asthma. *Clin Exp Allergy*, 46(6):793–802, 2016. doi: 10.1111/cea.12695. URL <https://www.ncbi.nlm.nih.gov/pubmed/26685004>.
- [158] Richard Serfozo. *Basics of Applied Stochastic Processes, Springer 2009*. jan 2009. doi: 10.1007/978-3-540-89332-5.
- [159] Fatemeh Seyednasrollah, Devin C Koestler, Tao Wang, Stephen R Piccolo, Roberto Vega, Russell Greiner, Christiane Fuchs, Eyal Gofer, Luke Kumar, Russell D Wolfinger, Kimberly Kanigel Winner, Chris Bare, Elias Chaibub Neto, Thomas Yu, Liji Shen, Kald Abdallah, Thea Norman, Gustavo Stolovitzky, Howard R Soule, Christopher J Sweeney, Charles J Ryan, Howard I Scher, Oliver Sartor, Laura L Elo, Fang Liz Zhou, Justin Guinney, James C Costello, and Prostate

- Cancer DREAM Challenge Community. A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer. *JCO Clinical Cancer Informatics*, (1):1–15, 2017. doi: 10.1200/CCI.17.00018. URL <https://doi.org/10.1200/CCI.17.00018>.
- [160] T Sing, O Sander, N Beerenwinkel, and T Lengauer. ROCr: visualizing classifier performance in R. *Bioinformatics*, 21(20):7881, 2005. URL <http://rocr.bioinf.mpi-sb.mpg.de>.
- [161] Wacharasak Siriseriwan. *smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE*, 2016. URL <http://cran.r-project.org/package=smotefamily>.
- [162] Holt D Skinner C. J and T M F Smith. Analysis of complex surveys. 1989.
- [163] Marina Skurichina and Robert P W Duin. Bagging for linear classifiers. *Pattern Recognition*, 31:909–930, 1998.
- [164] Hon-cheong So, Johnny S H Kwan, Stacey S Cherny, and Pak C Sham. Risk Prediction of Complex Diseases from Family History and Known Susceptibility Loci , with Applications for Cancer Screening. *The American Journal of Human Genetics*, 88(5):548–565, 2011. ISSN 0002-9297. doi: 10.1016/j.ajhg.2011.04.001. URL <http://dx.doi.org/10.1016/j.ajhg.2011.04.001><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3146722/pdf/main.pdf>.
- [165] Athina Spiliopoulou, Reka Nagy, Mairead L Bermingham, Jennifer E Huffman, Caroline Hayward, Veronique Vitart, Igor Rudan, Harry Campbell, Alan F Wright, James F Wilson, Ricardo Pong-wong, Felix Agakov, Pau Navarro, and Chris S Haley. Genomic prediction of complex human traits : relatedness , trait architecture and predictive. 24(14):4167–4182, 2015. doi: 10.1093/hmg/ddv145. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4476450/pdf/ddv145.pdf>.
- [166] Ewout W. Steyerberg, Gerald J J M Borsboom, Hans C. van Houwelingen, Marinus J C Eijkemans, and J. Dik F Habbema. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Statistics in Medicine*, 23(16):2567–2586, 2004. ISSN 02776715. doi: 10.1002/sim.1844. URL http://onlinelibrary.wiley.com/store/10.1002/sim.1844/asset/1844_{_}ftp.pdf;jsessionid=DED31BF82BFBBC8803B6E40C17530A38.f04t03?v=1{&t=il2gqmvav{&s=e9fa09a4d91a5ad7818232e21535f74c19f85397.
- [167] R Core Team. R: A language and environment for statistical computing, 2016. URL <https://www.r-project.org/>.

- [168] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL <http://www.jstor.org/stable/2346178>.
- [169] Walter Torous and Rossen Valkanov. *Boundaries of Predictability: Noisy Predictive Regressions*. jan 2000.
- [170] M Trerotola, V Relli, P Simeone, and S Alberti. Epigenetic inheritance and the missing heritability. *Hum Genomics*, 9:17, 2015. doi: 10.1186/s40246-015-0041-3. URL <https://www.ncbi.nlm.nih.gov/pubmed/26216216>.
- [171] V Ullemar, P K E Magnusson, C Lundholm, A Zettergren, E Melén, P Lichtenstein, and C Almqvist. Heritability and confirmation of genetic association studies for childhood asthma in twins. *Allergy*, 71(2):230–238, 2016. ISSN 1398-9995. doi: 10.1111/all.12783. URL <http://doi.org/10.1111/all.12783>.
- [172] Eleonora P Uphoff, Philippa K Bird, Joseph Maria Antó, Mikel Basterrechea, Andrea Von Berg, Anna Bergström, and Jean Bousquet. Variations in the prevalence of childhood asthma and wheeze in MeDALL cohorts in Europe. *ERJ Open Research*, 2017. doi: 10.1183/23120541.00150-2016. URL <http://dx.doi.org/10.1183/23120541.00150-2016><http://openres.ersjournals.com/content/erjor/3/3/00150-2016.full.pdf>.
- [173] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(1):1–67, 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03. URL <https://www.jstatsoft.org/index.php/jss/article/view/v045i03>.
- [174] Joaquin Vanschoren, Jan N van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL <http://doi.acm.org/10.1145/2641190.2641198>.
- [175] M Vatin, S Bouvier, L Bellazi, X Montagutelli, P Laissue, A Ziyat, C Serres, P De Mazancourt, M N Dieudonne, E Mornet, D Vaiman, and J C Gris. Polymorphisms of human placental alkaline phosphatase are associated with in vitro fertilization success and recurrent pregnancy loss. *Am J Pathol*, 184(2):362–368, 2014. doi: 10.1016/j.ajpath.2013.10.024. URL <http://www.ncbi.nlm.nih.gov/pubmed/24296104>.
- [176] Ana I Vazquez, Yogasudha Veturi, Michael Behring, Sadeep Shrestha, Matias Kirst, Marcio F R Resende, and Gustavo de los Campos. Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use

- of Whole-Genome Multiomic Profiles. *Genetics*, 203(3):1425–1438, jul 2016. ISSN 0016-6731. doi: 10.1534/genetics.115.185181. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4937492/>.
- [177] D Vercelli. Gene-environment interactions in asthma and allergy: the end of the beginning? *Curr Opin Allergy Clin Immunol*, 10(2):145–148, 2010. doi: 10.1097/ACI.0b013e32833653d7. URL <http://www.ncbi.nlm.nih.gov/pubmed/20051845>.
- [178] Donata Vercelli. Discovering susceptibility genes for asthma and allergy. *Nature reviews. Immunology*, 8(3):169–82, mar 2008. ISSN 1474-1741. doi: 10.1038/nri2257. URL <http://www.ncbi.nlm.nih.gov/pubmed/18301422>.
- [179] W. H. DuMouchel and G. J. Duncan. Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*, 78(383):535–543, 1983. URL <http://www.stat.cmu.edu/~brian/905-2008/papers/DumouchelDuncan-JASA-1983.pdf>.
- [180] S W Wallace, A Magalhaes, and A Hall. The Rho target PRK2 regulates apical junction formation in human bronchial epithelial cells. *Mol Cell Biol*, 31(1):81–91, 2011. doi: 10.1128/MCB.01001-10. URL <http://www.ncbi.nlm.nih.gov/pubmed/20974804>.
- [181] S Weidinger, S A Willis-Owen, Y Kamatani, H Baurecht, N Morar, L Liang, P Edser, T Street, E Rodriguez, G M O’Regan, P Beattie, R Folster-Holst, A Franke, N Novak, C M Fahy, M C Winge, M Kabesch, T Illig, S Heath, C Soderhall, E Melen, G Pershagen, J Kere, M Bradley, A Lieden, M Nordenskjold, J I Harper, W H McLean, S J Brown, W O Cookson, G M Lathrop, A D Irvine, and M F Moffatt. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Hum Mol Genet*, 22(23):4841–4856, 2013. doi: 10.1093/hmg/ddt317. URL <https://www.ncbi.nlm.nih.gov/pubmed/23886662>.
- [182] S K Weiland, E von Mutius, T Hirsch, H Duhme, C Fritzsche, B Werner, A Husing, M Stender, H Renz, W Leupold, and U Keil. Prevalence of respiratory and atopic disorders among children in the East and West of Germany five years after unification. *Eur Respir J*, 14(4):862–870, 1999. URL <https://www.ncbi.nlm.nih.gov/pubmed/10573234>.
- [183] J E White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128, 1982. URL <http://aje.oxfordjournals.org/content/115/1/119.full.pdf>.

- [184] C Winkler, J Krumsiek, J Lempainen, P Achenbach, H Grallert, E Giannopoulou, M Bunk, F J Theis, E Bonifacio, and a G Ziegler. A strategy for combining minor genetic susceptibility genes to improve prediction of disease in type 1 diabetes. *Genes and immunity*, 13(7):549–55, oct 2012. ISSN 1476-5470. doi: 10.1038/gene.2012.36. URL <http://www.ncbi.nlm.nih.gov/pubmed/22932816>.
- [185] Christiane Winkler, Jan Krumsiek, Florian Buettner, Christof Angermüller, Eleni Z. Giannopoulou, Fabian J. Theis, Anette Gabriele Ziegler, and Ezio Bonifacio. Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes. *Diabetologia*, 57(12):2521–2529, 2014. ISSN 14320428. doi: 10.1007/s00125-014-3362-1. URL <http://download.springer.com/static/pdf/220/art-253A10.1007-252Fs00125-014-3362-1.pdf?originUrl=http-3A-2F-link.springer.com-2Farticle-2F10.1007-2Fs00125-014-3362-1-1&token2=exp=1482847409-~acl=-2Fstatic-2Fpdf-2F220-2Fart-25253A10.1007-25252Fs00125-014-3362>.
- [186] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, 10(3):515–534, jul 2009. ISSN 1465-4644. doi: 10.1093/biostatistics/kxp008. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2697346/>.
- [187] D H Wolpert and W G Macready. No Free Lunch Theorems for Optimization. *Trans. Evol. Comp*, 1(1):67–82, apr 1997. ISSN 1089-778X. doi: 10.1109/4235.585893. URL <https://doi.org/10.1109/4235.585893>.
- [188] Naomi Wray and Peter Visscher. Estimating Trait Heritability. *Nature Education*, 1(1):29, 2008.
- [189] Nr Wray, Me Goddard, and Pm Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17:1520–1528, 2007. doi: 10.1101/gr.6665407.1520. URL <http://genome.cshlp.org/content/17/10/1520.short>.
- [190] Marvin N Wright. *ranger: A Fast Implementation of Random Forests*, 2016. URL <http://cran.r-project.org/package=ranger>.
- [191] Jincao Wu, Ruth M Pfeiffer, and Mitchell H Gail. Strategies for developing prediction models from genome-wide association studies. *Genetic epidemiology*, 37(8):768–

- 777, 2013. doi: 10.1002/gepi.21762. URL <http://www.ncbi.nlm.nih.gov/pubmed/24166696>.
- [192] Chang Xu, Dacheng Tao, and Chao Xu. A Survey on Multi-view Learning. *Cvpr*, 36(8):300072, 2015. URL <http://arxiv.org/abs/1304.5634>.
- [193] B Zadrozny, J Langford, and N Abe. Cost-sensitive learning by cost-proportionate example weighting. *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 435–442, 2003. doi: 10.1109/ICDM.2003.1250950. URL <http://ieeexplore.ieee.org/xpls/abs/abs%7B%7Dall.jsp?arnumber=1250950>.
- [194] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. *Twenty-first international conference on Machine learning - ICML '04*, page 114, 2004. doi: 10.1145/1015330.1015425. URL <http://portal.acm.org/citation.cfm?doid=1015330.1015425>.
- [195] Qing Zhao, Xingjie Shi, Yang Xie, Jian Huang, BenChang Shia, and Shuangge Ma. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in Bioinformatics*, 16(2):291–303, mar 2015. ISSN 1467-5463. doi: 10.1093/bib/bbu003. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4375393/>.

Appendix A

Supplementary Material

A.1 On correcting classifiers for sample selection bias

A.1.1 Simulation scenarios with variables from different distribution families

In Chapter 3, Section 3.3 we have conducted a simulation study with many scenarios for different classifiers and for different distributions. In order to convey a bigger picture, we created a further more heterogeneous distribution scenario: We simulated variables from different distributions within one data set which are partly correlated and with an interaction effect on the outcome. We also added an additional noise variable (which was not known/included for the training process).

Design

Concretely we generated the data analogously to the other scenarios of Section 4.1 with several changes.

The variables were generated as follows:

- $\tilde{X}^{(1)} \sim \mathcal{N}(0, 1)$
- $\tilde{X}^{(2)} \sim t(25)$

- $\tilde{X}^{(3)} \sim \tilde{X}^{(1)} + \mathcal{N}(0, 0.36)$
- $\tilde{X}^{(4)} \sim \tilde{X}^{(2)} + \mathcal{N}(0, 1.69)$
- $\tilde{X}^{(5)} \sim \text{Ber}(0.6)$
- $\tilde{X}^{(6)} \sim \mathcal{N}(0, 1)$
- $\tilde{X}^{(7)} = \tilde{X}^{(1)} * \tilde{X}^{(5)}$

Into our models we included $\tilde{X}^{(j)}$ for $j = 1, \dots, 5$, so that $\tilde{X}^{(6)}$ represents noise for constructing Y and $\tilde{X}^{(7)}$ an interaction. The corresponding effects were chosen to be $\boldsymbol{\beta} = (\beta_e, \beta_1, \dots, \beta_7) = (0.5, 0.1, -0.12, 0.07, 0.05, -0.9, 0.07, 0.9)$.

Results

The performances for the simulation scenario for the four classifiers, logistic regression, random forest, logistic regression with interaction terms, and naive Bayes, are compared in Figure A.1: We fit a linear model for the AUC as influenced by the correction method (dummy-coded, no correction as reference category). The graphic depicts 95% confidence intervals for the respective coefficients. The dotted line shows the intercept of the model, i.e. the mean AUC for no correction. The blue colored methods are newly proposed by us.

A.1.2 Investigation of varying sample size and degree of imbalance

In Chapter 3 we have identified the parametric inverse-probability bagging as the most successful correction approach for the random forest. In this section we investigate if this also holds when diverse parameters in the data are changed. Therefore we compare learning on population, non-correcting, IP oversampling as one standard correction method and parametric IP bagging.

For simulating the data we used the same settings as in Section A.1.1; however, we varied sample size of the data taken from the population and in a further scenario varied the distributions of the stratum variables Y and X_e , i.e. made rare categories less or more rare.

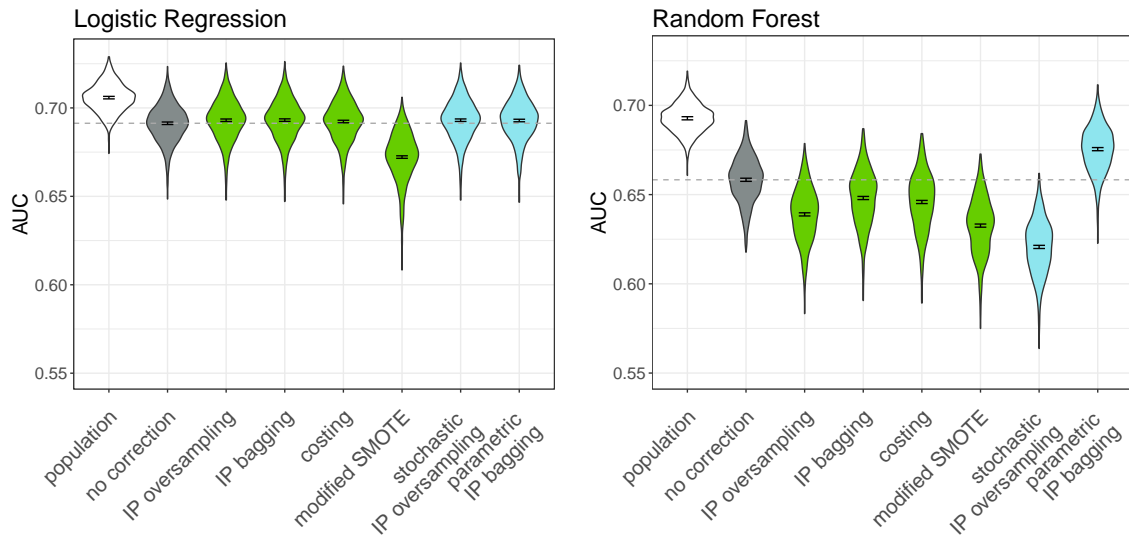


Figure A.1a: Performance of correction approaches for mixed distributed features for logistic regression, measured by AUC. All approaches perform significantly better than no correction except for the modified SMOTE approach.

Figure A.1b: Performance of correction approaches for mixed distributed features for random forest, measured by AUC. Only parametric IP bagging performs significantly better than no correction.

Conducting the simulation study for different sample sizes led to poor performance for correction approaches when the sample size was smaller than the original one of $n = 500$. When the sample size was increased the parametric IP bagging worked well (Figure A.2).

Varying the imbalance of the Y and X_e in the population, i.e. the original probabilities $P(Y = 1) = 0.1$ and $P(X_e = 1) = 0.1$ show that our approach still outperforms the non-correction approach and the IP oversampling when the rare categories are less rare, but not when the rareness is more extreme (Figure A.3). However, this may be explained by the results for varying sample sizes: Due to a restriction of the sampling design we had to decrease to sample size for the the more extreme case from $n = 500$ to $n = 200$ in this scenario; as observed above (Figure A.2) this sample size is too small in general for the approach to be effective.

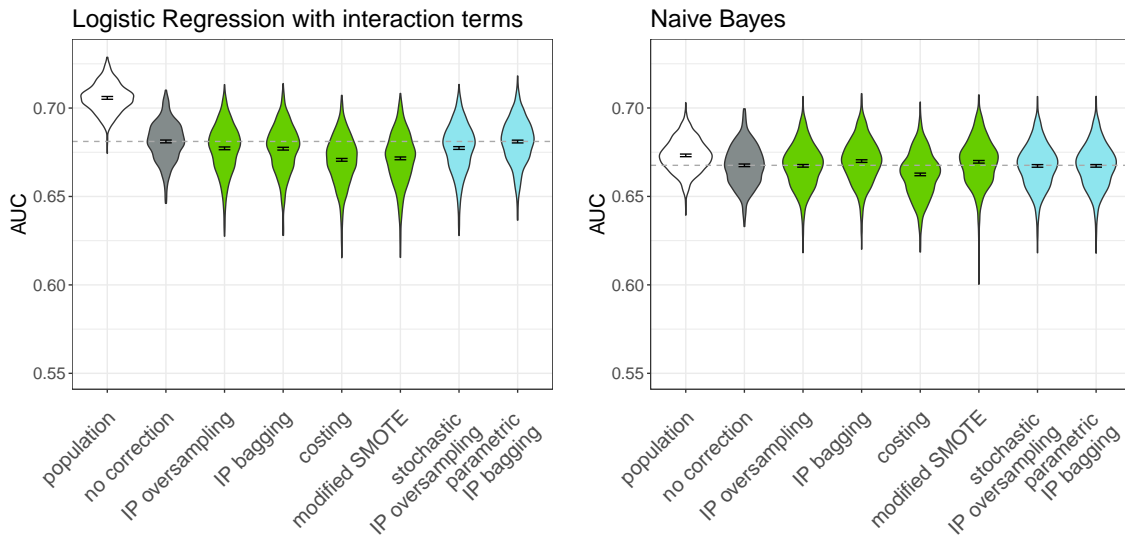


Figure A.1c: Performance of correction approaches for mixed distributed features for logistic regression with interaction effects, measured by AUC. All approaches perform significantly worse than no correction except parametric IP bagging which is not significantly different.

Figure A.1d: Performance of correction approaches for mixed distributed features for naive Bayes, measured by AUC. Only IP bagging and modified SMOTE perform significantly better than no correction.

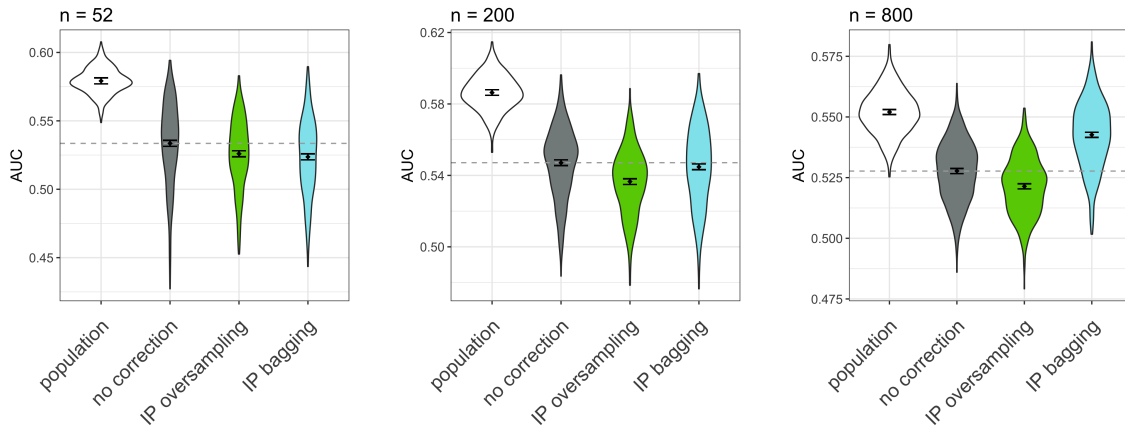


Figure A.2: Performance of correction approaches for mixed distributed features for random forest, for different sample sizes. Both, parametric IP oversampling and parametric IP bagging fail to outperform no correction when the sample size is small (left and center figure). Parametric IP bagging outperforms no correction for a bigger sample size (right figure).

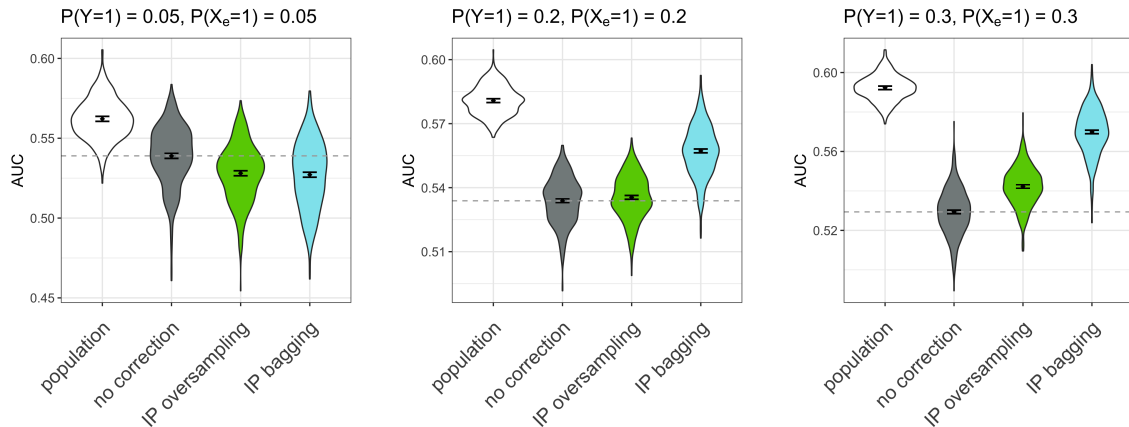


Figure A.3: Performance of correction approaches for mixed distributed features for random forest, for different degrees of imbalance of Y and X_e . Both, parametric IP oversampling and parametric IP bagging fail to outperform no correction when $P(Y = 1)$ and $P(X = 1)$ are both small (left figure), but parametric IP bagging outperforms no correction when the corresponding probabilities are increased (center and right figure).

A.1.3 Investigation of other bias types

So far, in our studies on sample selection bias we had focused on complete bias which is a combination of the other two types of bias — feature bias and label bias. Therefore, we explored the behavior of parametric IP bagging for the random forest in this case as for this classifier only our approach had successfully corrected for sample selection bias.

As in the previous section we applied the same settings as in Section A.1.1 for generating the data, but used only a one-phase procedure as sampling design in order to generate bias types where either only a rare outcome variable or only a rare feature was enriched (label and feature bias, respectively).

The results clearly show that for both types of bias, for label bias (Figure A.4) and for feature bias (Figure A.5) the parametric IP bagging outperforms the non-correction approach. IP oversampling, however, performs less well than not correcting.

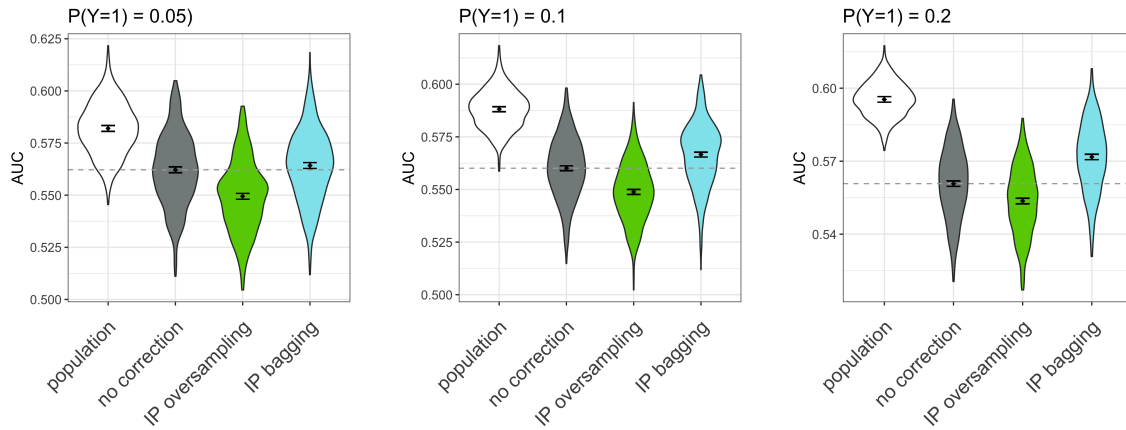


Figure A.4: Performance of correction approaches for mixed distributed features for random forest, for label bias and varying probabilities $P(Y = 1)$. Parametric IP bagging outperforms no correction in all cases.

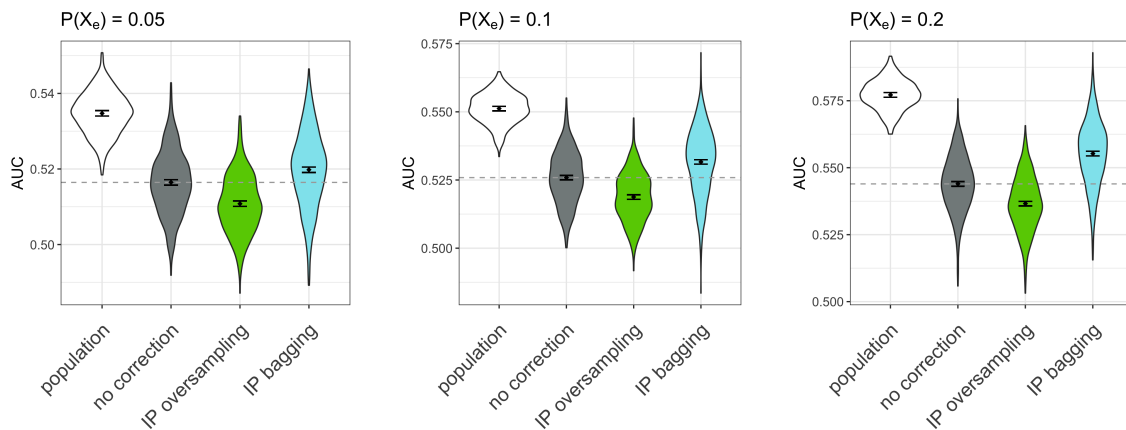


Figure A.5: Performance of correction approaches for mixed distributed features for random forest, for feature bias and varying probabilities $P(X_e = 1)$. Parametric IP bagging outperforms no correction in all cases.

A.2 On analyses on GABRIELA study

A.2.1 P-value adjustment for variable importance

Background

In the frame of Chapter 4 we did further investigations on p-value adjustments. As in the chapter we did not only present some measure for the importance of variables in a classification model but also delivered p-values for more than 20 variables, so that multiple testing may be an issue.

In practice, when a generalized linear regression model is applied and presented with all its significance tests and p-values, the multiple testing issue often is not taken into account, as usually the number of covariates is small so that false rejections of the null hypotheses are unlikely. However, p-value adjustment can be done in such a scenario [125].

We applied a less commonly used method to detect significant variables in Chapter 4 — the non-parametric Altmann approach for random forests [4] and want to clarify if p-value adjustment is necessary here. As we have up to 40 variables at hand, we perform a simulation study with this many variables and a sufficient amount of observations. We compare it to other methods: the generalized linear regression model, the LASSO (in terms of how many variables are estimated to non-zero), and another approach for random forest proposed by Janitza et al. [89].

We set up a simulation scenario for a binary outcome variable and normally and Bernoulli distributed variables (varying n and p) where we applied the four methods. 5 variables always had truly significant influence on the outcome.

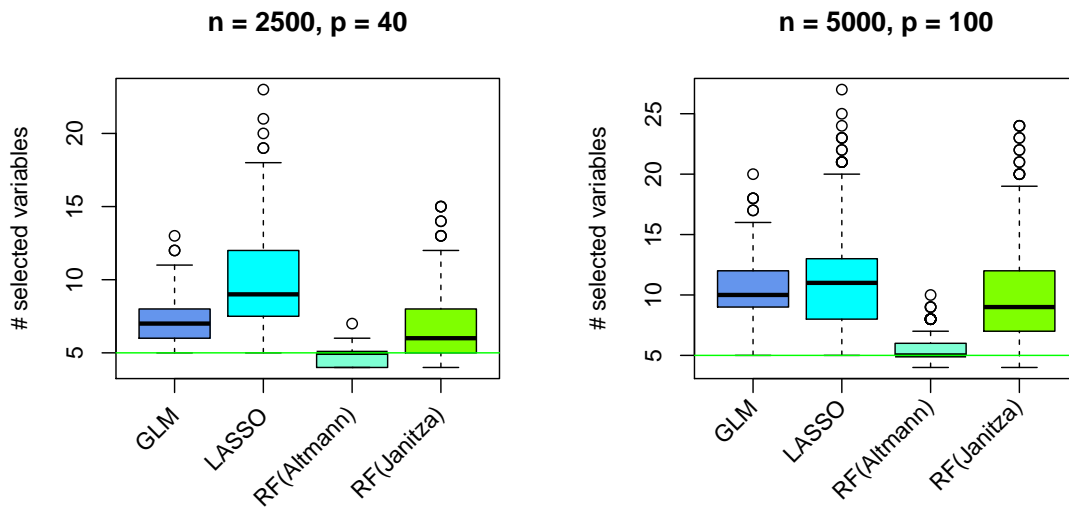


Figure A.6a: Boxplots for the number of variables that were selected as significant by the model for $p = 40$ and $n = 2500$. 5 variables have had a true effect. Only the Altmann approach did not exceed the number of significant variables.

Figure A.6b: Boxplots for the number of variables that were selected as significant by the model for $p = 100$ and $n = 5000$. 5 variables have had a true effect. Only the Altmann approach did not exceed the number of significant variables.

Results

We show the results for the four different variable importance measures in terms of type I and type II errors. Figures A.6a and A.6b show how many variables were selected as significant variables when 5 variables truly had an actual effect.

Type I error. Even though it less occurred in Janitzza et al. [89] for the non-parametric Altmann approach, in our simulation study the type I error was much smaller than 0.05 (Figures A.6c and A.6d). This held only for this approach and was 0.05 for the GLM and the approach of Janitzza. LASSO had a slightly higher type I error, given that selected variables are seen as significant variables.

Type II error. There was no type II error for GLM or LASSO, and a type II error of 20% for the two random forest approaches (Figures A.6e and A.6f). This, however, still corresponds to a statistical power of 80%.

Note that for these settings the number of variables was too little to guarantee stable results for the method of Janitzza et al. [89].

For the project of Chapter 4 we can conclude that the non-parametric Altmann approach

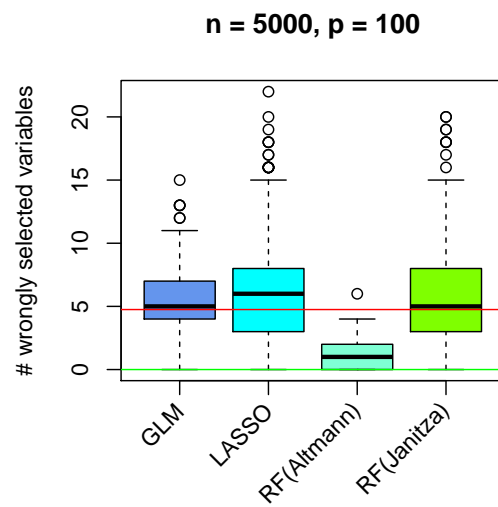
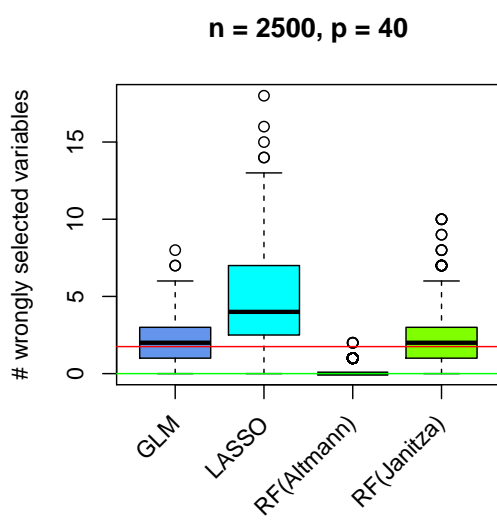


Figure A.6c: Boxplots for the number of variables that were wrongly selected as significant by the model for $p = 40$ and $n = 2500$. The false positive rate was 0 (green line) only for the Altmann approach. The median of all other approaches was at least at 5% (red line) for all other approaches.

Figure A.6d: Boxplots for the number of variables that were wrongly selected as significant by the model for $p = 100$ and $n = 5000$. The false positive rate was 0 (green line) for lower quantile of the Altmann approach (median: 1%). The median of all other approaches was at least at 5% (red line) for all other approaches.

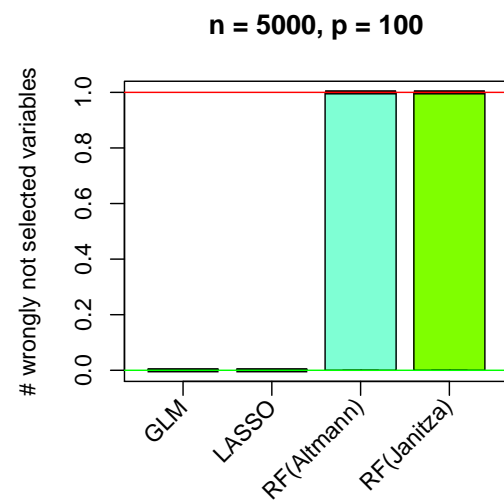
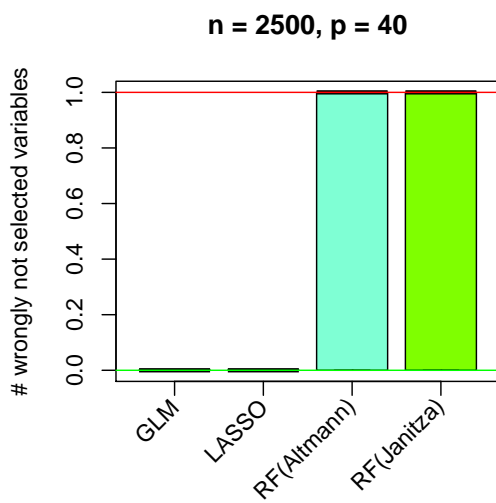


Figure A.6e: Boxplots for the number of variables that were wrongly not selected as significant by the model for $p = 40$ and $n = 2500$. The false negative rate was 0 (green line) only for the GLM and the LASSO. The median of the random forest approaches was at least at 20% (red line) for all other approaches which still corresponds to a power of 80 %.

Figure A.6f: Boxplots for the number of variables that were wrongly not selected as significant by the model for $p = 100$ and $n = 5000$. The false negative rate was 0 (green line) only for the GLM and the LASSO. The median of the random forest approaches was at least at 20% (red line) for all other approaches which still corresponds to a power of 80 %.

rather failed to detect a small percentage of truly significant variables, but did not erroneously determine variables as significant when there was no true effect.

A.2.2 Variable importance genome-wide

In the main chapter we determined the most important variables for the final best model. We did so for the best model among the genome-wide models. Using different groups of variables separately yielded only one AUC significantly different from 0.5 — the random forest for farm children (Figure 4.3). We assessed the variable importance for this model (Figure A.7A) and for the same constellation when family history and demographics are added (Figure A.7B).

For genome-wide variable importance with farm-exposure as outcome variable calculated for asthmatics and non-asthmatics (cf. Chapter 4), generally more significant variables were found for asthmatics than for non-asthmatics for different levels of significance (Figure A.8).

A.2.3 Tables on family history, environment and genetics

Supplementary to the analyses of the main chapter we provide detailed information on variables (family history, demographics, environment) of the GABRIELA study (Table A.1) and the PASTURE study (Table A.2). Table A.3 shows information on the candidate SNPs extracted from other GWAS.

Table A.1: Distribution of environmental determinants and family history of atopy in GABRIELA

characteristic	cases	controls	p-value
center (study center (ref.: Innsbruck))	16.4% (Basel) 27.8% (Munich) 37.7% (Ulm)	15.6% (Basel) 32.6% (Munich) 31.9% (Ulm)	0.123
female (female gender)	39.70%	49.40%	0.002
age (age in years at 2007-01-01)	8.32 (se=0.06)	8.19 (se=0.06)	0.15
BMI (body mass index)	17.11 (se=0.11)	16.99 (se=0.11)	0.375
farm (farming status)	9%	13.60%	<0.001
siblings (>1 siblings)	41.80%	44.60%	0.374
parental-education (high parental education)	27.30%	28.80%	0.633
smoking-pregnancy (maternal smoking in pregnancy)	12.40%	8.50%	0.037
milk-last-yr (consumption of farm milk past 12 months)	13.40%	19.40%	<0.001
milk-first-yr (consumption of farm milk in first year of life)	6.20%	11.80%	<0.001
milk-first-3yrs (consumption of farm milk pregnancy to age 3yrs)	20.70%	27.60%	<0.001
cow-last-yr (contact with cows past 12 months)	12.90%	16.60%	0.02
cow-first-3yrs (contact with cows pregnancy to age 3yrs)	14.60%	20.30%	0.001
straw-last-yr (contact with straw past 12 months)	15.70%	21.10%	0.009
straw-first-3yrs (contact with straw pregnancy to age 3yrs)	12.40%	16.20%	0.009
hay-last-yr (contact with hay past 12 months)	29.70%	33.50%	0.145
hay-first-3yrs (contact with hay pregnancy to age 3yrs)	21.60%	26.10%	0.028
cow/straw-first-3yrs	5.1% (straw only) 7.5% (cow only) 7.2% (both)	4.8% (straw only) 8.6% (cow only) 11.4% (both)	0.024
(contact with cows and/or straw pregnancy to age 3yrs)	8.7% (straw only)	11.9% (straw only)	
cow/straw-last-yr (contact with cows and/or straw past 12 months)	5.9% (cow only) 7.1% (both)	7.1% (cow only) 9.4% (both)	0.029
barn-first-3yrs (stay in barn pregnancy to age 3yrs)	14.60%	18.50%	0.015
barn-last-yr (stay in barn past 12 months)	16.50%	20.70%	0.021
stable-weekly-last-yr (stay in cattle stable once/week)	8.70%	12%	0.016
traffic (how often do trucks or busses drive by)	35.9% (rarely) 27.6% (often during day) 9.5% (almost whole day)	39.3% (rarely) 26.9% (often during day) 10.1% (almost whole day)	0.558
mold-last-yr (rooms with visible mold)	18%	11.60%	0.004
dog-first-yr (dog allowed to stay in the room in first year of life)	5.30%	6.50%	0.456
cat-first-yr (cat allowed to stay in the room in first year of life)	8.50%	15.50%	0.001
dog-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)	6.90%	7.20%	0.838
cat-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)	11.50%	21.40%	<0.001
dog-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)	6.30%	6.90%	0.736
cat-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)	15%	24.60%	<0.001
dog-last-yr (dog allowed to stay in the room past 12 months)	7.70%	8.90%	0.506
cat-last-yr (cat allowed to stay in the room past 12 months)	19.80%	30.10%	<0.001
children-household	2.39 (se=0.04)	2.37 (se=0.04)	0.921
(people between 0-18 years live currently in your household)			
adults-household	2.13 (se=0.06)	2.18 (se=0.03)	0.005
(people 18 years and older live currently in your household)			
parents-smoke-ever (have parents ever smoked after birth)	61.60%	55.60%	0.052
parents-smoke-currently (do parents currently smoke)	33.20%	26.80%	0.024
daycare			
(regular attendance of facilities with children until school enrolment)	92%	92%	0.981
antibiotics-pregnancy (mother takes antibiotics during pregnancy)	11.80%	8.30%	0.074
cattle-last-yr (stay in cattle stable past 12 months)	11.10%	14.20%	0.043
cattle-first-yr (stay in cattle stable in first year of life)	6.10%	10.20%	<0.001
birth-season (season of birth (ref.: Summer))	25.6% (Spring) 27.2% (Autumn) 25.4% (Winter)	22.6% (Spring) 23.5% (Autumn) 25.6% (Winter)	0.11
family-atopy (parental atopy or sibling atopy)	70%	49.70%	<0.001
family-asthma (parental asthma or sibling asthma)	30.60%	12.40%	<0.001
family-hayfev (parental hay fever or sibling hay fever)	48.30%	36.60%	<0.001
family-fheczema (parental eczema or sibling eczema)	36.50%	25.80%	<0.001

* p-values based on Fisher's exact test or, in case of continuous variables, Wilcoxon tests
se = standard error of mean, yr = year

Table A.2: Distribution of environmental determinants and family history of atopy in PASTURE

characteristic	cases	controls	p-value
female (female gender)	33.30%	50.10%	0.005
age (age in years at 2007-01-01)	6.12 (se=0.03)	6.12 (se=0.01)	0.823
BMI (body mass index)	15.73 (se=0.17)	15.78 (se=0.07)	0.501
farm (farming status)	37.20%	49.30%	0.041
siblings (>1 siblings)	33.30%	34%	0.905
parental-education (high parental education)	89.70%	82.90%	0.121
smoking-pregnancy (maternal smoking in pregnancy)	14.50%	12.10%	0.555
milk-last-yr (consumption of farm milk past 12 months)	30.80%	43.20%	0.034
milk-first-yr (consumption of farm milk in first year of life)	21.80%	33.50%	0.035
milk-first-3yrs (consumption of farm milk pregnancy to age 3yrs)	41.10%	53.80%	0.038
cow-last-yr (contact with cows past 12 months)	23.70%	40.50%	0.004
cow-first-3yrs (contact with cows pregnancy to age 3yrs)	31.40%	44.50%	0.035
straw-last-yr (contact with straw past 12 months)	19.70%	40%	<0.001
straw-first-3yrs (contact with straw pregnancy to age 3yrs)	26.90%	43.80%	0.007
hay-last-yr (contact with hay past 12 months)	25.60%	42.80%	0.003
hay-first-3yrs (contact with hay pregnancy to age 3yrs)	38.20%	46.70%	0.181
cow/straw-first-3yrs	0% (straw only)	2.1% (straw only)	
(contact with cows and/or straw pregnancy to age 3yrs)	3% (cow only)	3.2% (cow only)	0.053
	27.3% (both)	41.7% (both)	
	1.4% (straw only)	5.1% (straw only)	
cow/straw-last-yr (contact with cows and/or straw past 12 months)	6.8% (cow only)	6.7% (cow only)	0.006
	17.6% (both)	34% (both)	
barn-first-3yrs (stay in barn pregnancy to age 3yrs)	84%	77.60%	0.457
barn-last-yr (stay in barn past 12 months)	20.50%	42.80%	<0.001
stable-weekly-last-yr (stay in cattle stable once/week)	29.50%	52.30%	<0.001
mold-last-yr (rooms with visible mold)	15.40%	16%	0.891
dog-first-yr (dog allowed to stay in the room in first year of life)	15.40%	18.70%	0.472
cat-first-yr (cat allowed to stay in the room in first year of life)	26.90%	29.30%	0.658
dog-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)	18.70%	17.10%	0.726
cat-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)	25.70%	30.70%	0.367
dog-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)	25.30%	33.10%	0.168
cat-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)	37.80%	54.90%	0.005
dog-last-yr (dog allowed to stay in the room past 12 months)	16.70%	19.10%	0.602
cat-last-yr (cat allowed to stay in the room past 12 months)	34.60%	40.60%	0.299
children-household			
(people between 0-18 years live currently in your household)	1.26 (se=0.13)	1.16 (se=0.04)	0.395
adults-household			
(people 18 years and older live currently in your household)	2.05 (se=0.03)	2.18 (se=0.02)	0.106
parents-smoke-ever (have parents ever smoked after birth)	68.90%	61.30%	0.198
parents-smoke-currently (do parents currently smoke)	29.70%	27.10%	0.621
daycare			
(regular attendance of facilities with children until school enrolment)	93.80%	92.30%	0.682
antibiotics-pregnancy (mother takes antibiotics during pregnancy)	29.90%	25.80%	0.441
cattle-last-yr (stay in cattle stable past 12 months)	35.30%	45.40%	0.109
cattle-first-yr (stay in cattle stable in first year of life)	35.10%	47.20%	0.042
	25.6% (Spring)	27% (Spring)	
birth-season (season of birth (ref.: Summer))	23.1% (Autumn)	22.8% (Autumn)	0.642
	32.1% (Winter)	26.2% (Winter)	
family-atopy (parental atopy or sibling atopy)	83.10%	61.10%	<0.001
family-asthma (parental asthma or sibling asthma)	35.10%	14.20%	<0.001
family-hayfev (parental hay fever or sibling hay fever)	68.90%	44.20%	<0.001
family-fheczema (parental eczema or sibling eczema)	50%	30.20%	<0.001

* p-values based on Fisher's exact test or, in case of continuous variables, Wilcoxon tests
se = standard error of mean, yr = year

Table A.3: SNPs associated with childhood asthma from GWAS Catalog. SNPs associated with childhood asthma in other studies are given by the GWAS Catalog. The 19 SNPs given in GABRIELA and used for analysis are highlighted in green.

SNP	Region	Location	Reported Genes	Mapped genes	Study
rs17036023-?	1p13.1	chr1:116587089	IGSF3	IGSF3	Ding L (PMID: 23829686); 2013
rs7527074-?	1q25.3	chr1:180676305	XPR1	XPR1	Ding L (PMID: 23829686); 2013
rs4658627-A	1q44	chr1:244347874	C1orf100	ZBTB18 - C1orf100	Forno E (PMID: 22560479); 2012
rs6054973-?	20p12.3	chr20:7405311	intergenic	MIR8062 - SRSF10P2	Ding L (PMID: 23829686); 2013
rs6721181-?	2p22.1	chr2:39888556	intergenic	THUMP2 - SLC8A1-AS1	Ding L (PMID: 23829686); 2013
rs17033506-?	3p22.3	chr3:35598334	intergenic	LOC100130503	Ding L (PMID: 23829686); 2013
rs9815663-T	3p26.2	chr3:3573203	IL5RA	CRBN - SUMF1	Forno E (PMID: 22560479); 2012
rs9823506-?	3q12.2	chr3:100757869	ABI3BP		
rs2705520-?	3q13.2	chr3:112550440	ATG3	ATG3	Ding L (PMID: 23829686); 2013
rs9883878-?	3q26.32	chr3:178137844	intergenic	FGFR3P4 - LINC01014	Ding L (PMID: 23829686); 2013
rs35141484-?	4p14	chr4:39086721	KLHL5	KLHL5	Ding L (PMID: 23829686); 2013
rs17218161-?	4q12	chr4:58347679	intergenic	SRIP1 - MIR548AG1	Ding L (PMID: 23829686); 2013
rs6871536-C	5q31.1	chr5:132634182	RAD50	RAD50	Bonnelykke K (PMID: 24241537); 2013
rs1295686-T	5q31.1	chr5:132660151	IL13	IL13	Bonnelykke K (PMID: 24241537); 2013
rs7770848-?	6p21.1	chr6:44801500	intergenic	N/A	Ding L (PMID: 23829686); 2013
rs2473967-?	6q21	chr6:113158133	intergenic	PA2G4P5 - SOCS5P5	Ding L (PMID: 23829686); 2013
rs886448-?	7p15.3	chr7:24200546	intergenic	RNA5SP228 - NPY	Ding L (PMID: 23829686); 2013
rs6967330-A	7q22.3	chr7:106018005	CDHR3	CDHR3	Bonnelykke K (PMID: 24241537); 2013
rs7807274-?	7q32.3	chr7:131336340	MKLN1	MKLN1	Ding L (PMID: 23829686); 2013
rs9297216-?	8p12	chr8:34187743	intergenic	CYCSP3 - RPL10AP3	Ding L (PMID: 23829686); 2013
rs16929097-?	9p23	chr9:12521826	intergenic	JKAMPP1 - TYRP1	Ding L (PMID: 23829686); 2013
rs928413-G	9p24.1	chr9:6213387	IL33	IL33	Bonnelykke K (PMID: 24241537); 2013
rs11141597-?	9q21.33	chr9:86912543	intergenic	RPS6P13 - GAS1	Ding L (PMID: 23829686); 2013
rs11000019-?	10q22.1	chr10:71831773	PSAP	PSAP	Ding L (PMID: 23829686); 2013
rs12570188-?	10q24.2	chr10:99095945	HPSE2	HPSE2	Ding L (PMID: 23829686); 2013
rs7927044-A	11q24.2	chr11:127891771	NR	KIRREL3-AS3 - ETS1	Forno E (PMID: 22560479); 2012
rs7328278-C	13q13.3	chr13:35777629	NR	DCLK1	Forno E (PMID: 22560479); 2012
rs10521233-G	17p12	chr17:13655763	NR	MIR548H3 - CDRT15P1	Forno E (PMID: 22560479); 2012
rs2305480-G	17q12 / 17q21.1	chr17:39905943	GSDMB	GSDMB	Bonnelykke K (PMID: 24241537); 2013
rs7216389-T	17q12 / 17q21.1	chr17:39913696	ORMDL3	GSDMB	Moffatt MF (PMID: 17611496); 2007
rs3894194-A	17q21.1	chr17:39965740	GSDMA	GSDMA	Bonnelykke K (PMID: 24241537); 2013

A.2.4 External validation for non-farm and farm children

In this section we report results for performance of a final prediction model determined on only non-farm children and only farm children validated on the Austrian arm of GABRIELA. For non-farm children random forest using demographics, family history and environment had performed best and thus was applied as a final model. This yielded an AUC of 0.63 (Figure A.9A.)

On farm children IPF-LASSO and random forest had performed similarly using demographics, family history and SNPs. We performed an average model combining prediction scores of both models: the prediction scores of the random forest was standardized with respect to mean and standard deviation of the IPF-LASSO's prediction score. The average of the two scores was used as final score and led to an AUC of 0.86 (Figure A.9B). Here, for guaranteeing robustness of AUC confidence intervals, high class imbalance (occurring for farm children on the Austrian GABRIELA arm) was avoided by forcing at least 10% asthma cases into each bootstrap sample. Figure A.10 indicates that the more cases are forced into one bootstrap sample the more precise the estimation of the AUC.

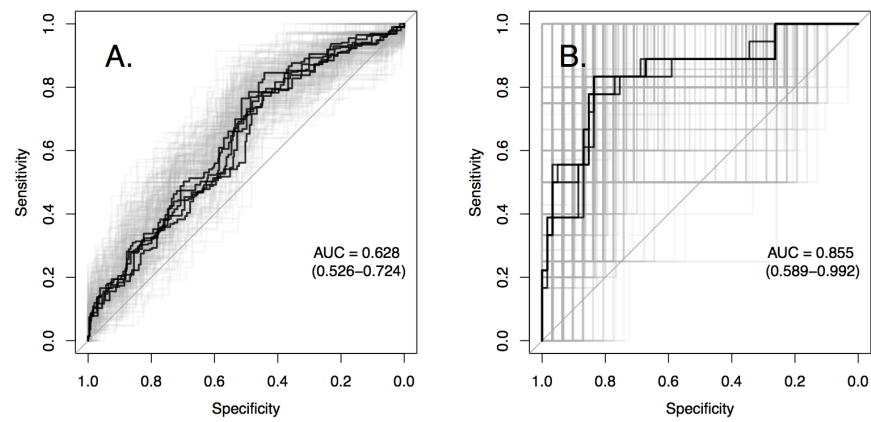


Figure A.9: ROC-curves for best models for non-farm and farm children validated on the Austrian arm of GABRIELA. Mean weighted ROC-curves of the 5 imputations (black) with 1000 weighted ROC curves obtained via bootstrap (gray) for predictions on the Austrian GABRIELA arm. A. Best prediction model for non-farm children (random forest) validated on the Austrian GABRIELA arm. B. Best prediction model, resulting from model averaging over standardized (subtraction of mean, division by standard deviation) prediction scores of random forest and IPF-LASSO, for farm children.

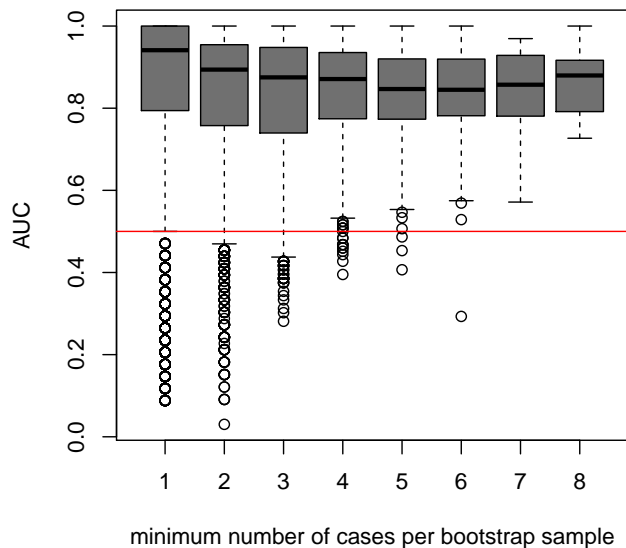


Figure A.10: Boxplots for AUCs for farm children validated on the Austrian arm of GABRIELA for different numbers of cases that were included into a bootstrap sample at minimum. 5000 bootstrap samples were taken for each boxplot. The boxplots' range decreases the more balanced the bootstrap samples are regarding their outcome.

A.2.5 Parametric inverse-probability bagging on GABRIELA

The novel approach parametric IP bagging we proposed in Chapter 3 was for completion applied on the data of the GABRIELA study. We applied the same variable setting as for the final random forest models in Chapter 4 which were validated on the Austrian GABRIELA arm. The results are shown in Figure A.11 and partly could be expected according to the investigations in Chapter 3: for all children and non-farm children mostly binary variables are given. This makes parametric IP bagging as it was proposed not directly suitable but simulation study results of Chapter 3 have shown that in such cases it still may work. The AUCs in our results were indeed only slightly lower than the ones for the standard random forest. Farm children, however, contained mostly SNPs. Those, seen as continuous variables, usually follow a non-symmetric distribution and thus may disturb the estimation of a multivariate normal distribution. Indeed, the approach results in an AUC close to 0.5.

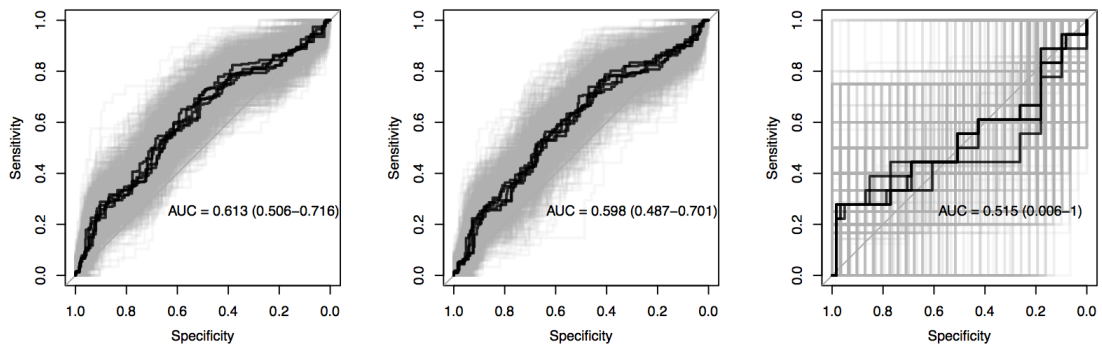


Figure A.11: External validation of parametric IP bagging on the Austrian GABRIELA arm using family history and demographics for learning. Left figure: for all children using environment in addition. Center figure: for non-farm children using environment in addition. Right figure: for farm children using candidate SNPs in addition.

A.3 On analyses on CLARA study

A.3.1 Wider selection of important variables

In a more relaxed selection we regarded all variables which were selected by LASSO and at the same time substantially correlated (correlation coefficient > 0.6) with a top-ranked variable identified by boosting. Figure A.12 shows the corresponding correlation structure. Figure A.13 illustrates the overlaps of correlated variables between variables identified by LASSO und boosting with the relaxed assumptions. Table A.4 gives the corresponding variable names.

Table A.4: Overlap of variables selected by LASSO and variables ranked among the top 50 by boosting with treating variables of the two algorithms as equal if their Pearson correlation coefficient was greater or equal 0.6. Variables in bold letters are contained in the LASSO selected and the boosting selected variables.

features LASSO	features LASSO correlated to features boosting	features boosting correlated to features boosting	features boosting
CD93 (LpA)	ALPP (CD328)	C1orf213//ZNF436 (CD328)	RORC (CD328)
CAPN13 (M)	SP9 (CD328)	TMEM106C (CD328)	OR51Q1 (CD328)
LOC100291105//RBM38 (M)	PKN2 (LpA)	ALPP (CD328)	S100A7 (CD328)
LINC00969 (M)	ALG5 (LpA)	GOLGA6L6//GOLGA6L1 (CD328)	IL1beta_M
LOC728613 (M)	C15orf59 (LpA)	FANK1 (CD328)	CCDC185 (CD328)
	CYTH1 (LpA)	CLCNKB (CD328)	Leukozyten
	ZNF432 (LpA)	NT5C3B (CD328)	IL5_CD328
	OR6B3 (LpA)	PARK7 (CD328)	STMN1 (CD328)
	IQGAP2 (LpA)	RD3 (LpA)	LURAP1//POMGNT1 (CD328)
	CSNK1G3 (LpA)	UQCRHL//UQCRH (CD328)	IL5_LpA
	PTK2 (LpA)	PKN2 (LpA)	ARTN (CD328)
	MPLKIP (LpA)	PRADC1 (CD328)	months-breastfeeding
	FMR1 (LpA)	CD3D (CD328)	NUDT13 (CD328)
	ANKS4B (M)	TRAPPC3 (CD328)	OR10X1 (CD328)
	NPIP9//NPIP6 (M)	PTK2 (LpA)	CHCHD1 (CD328)
	EPHA6 (M)	SEPW1 (CD328)	SMYD2 (CD328)
	SEMA3A (M)		RPL11 (CD328)
			Segmented granulocytes
			CAP1 (CD328)
			DMBX1 (CD328)
			ZFP69B (CD328)
			CSRP3 (CD328)
			alpha1Antitrypsin
			MMP13 (CD328)
			NGF (CD328)
			RASGRF1 (CD328)
			IRF8_328
			TGM1 (CD328)
			RNPEP (CD328)
			RHEB (LpA)
			TNN (CD328)
			SSU72 (CD328)
			IL1beta_LpA
			RAB25 (CD328)

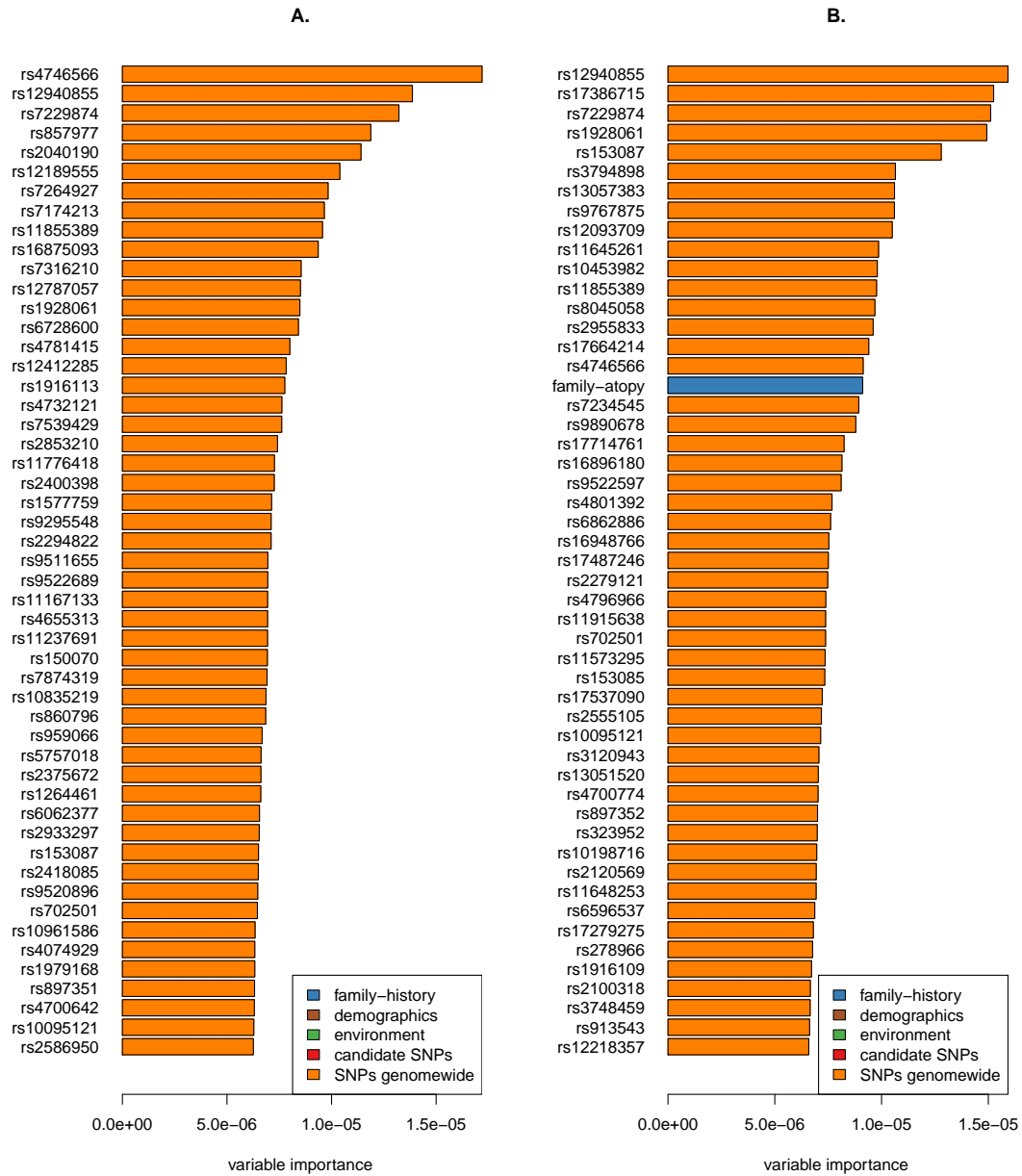


Figure A.7: Variable importance of the random forest model trained on genome-wide SNPs in farm children. A selection of the 50 most important variables is shown. A. Only the genome-wide SNPs are used for learning; corresponding to the lower right panel of Figure 4.3. B. Family history, demographics in addition to the genome-wide SNPs are used for learning; corresponding to the lower right panel of Figure 4.4.

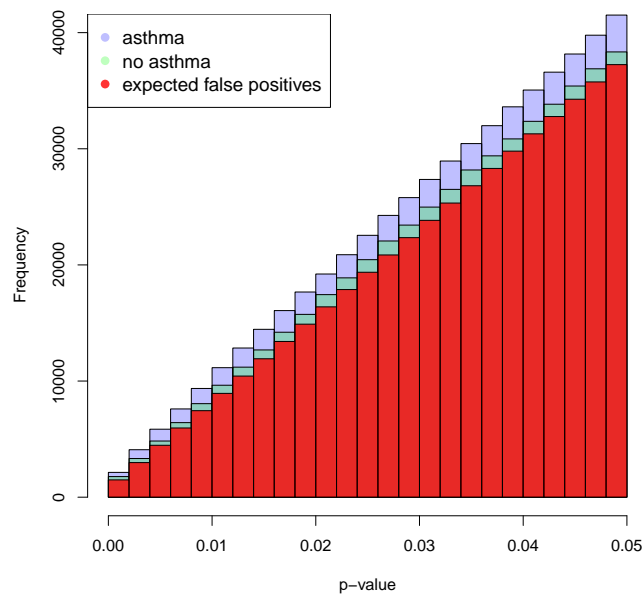


Figure A.8: Genome-wide association study of farming stratified for asthma cases and controls. The cumulative frequency of p-values is given stratified for asthma cases and healthy controls. The number of p-values representing false positive findings for the corresponding levels of significance is given as reference.

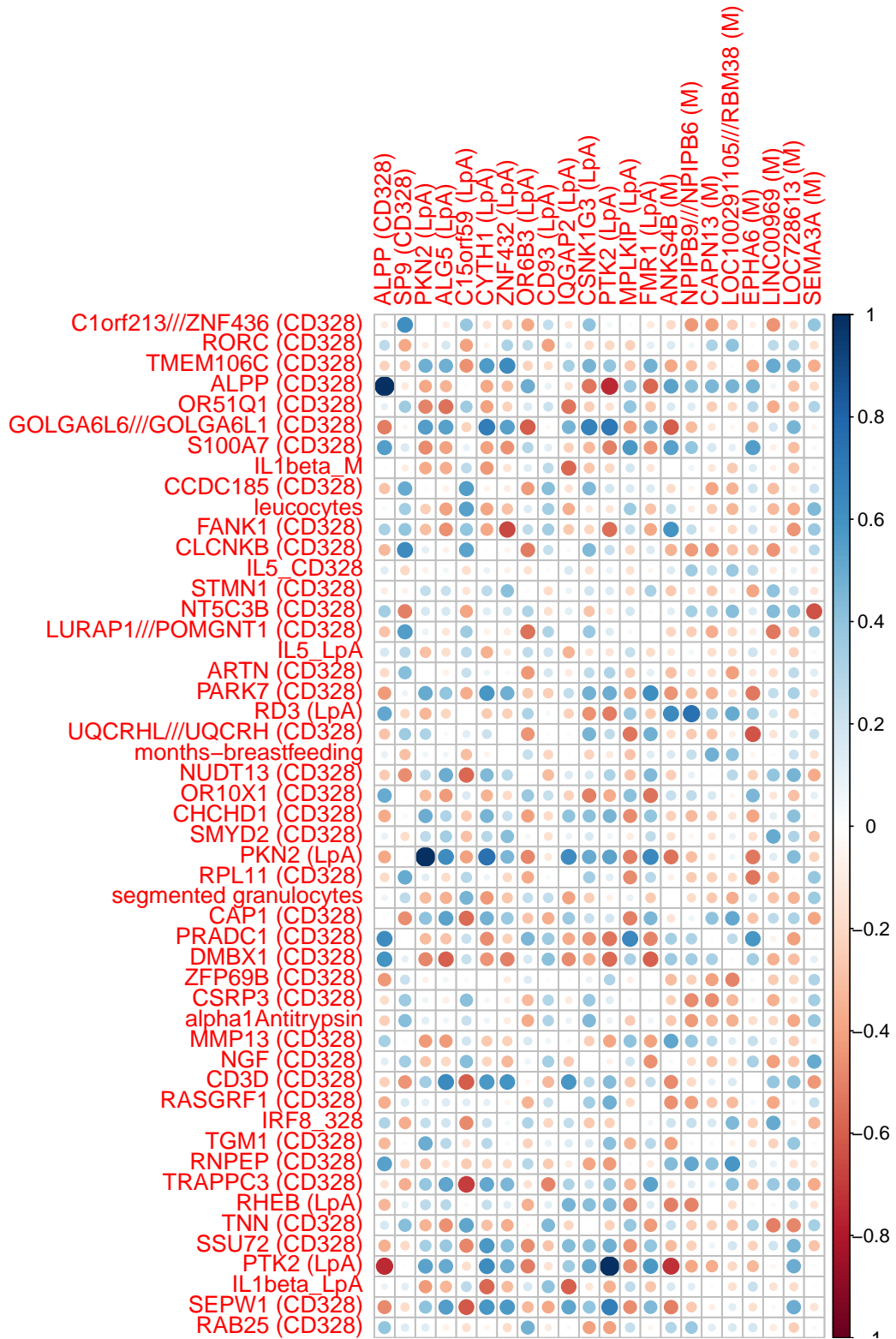


Figure A.12: Pearson correlation coefficient between variables selected by LASSO (columns) and those ranked under the top 50 by boosting (rows) in the complete case model. Genes are denoted by their names with the type of stimulation in parentheses.

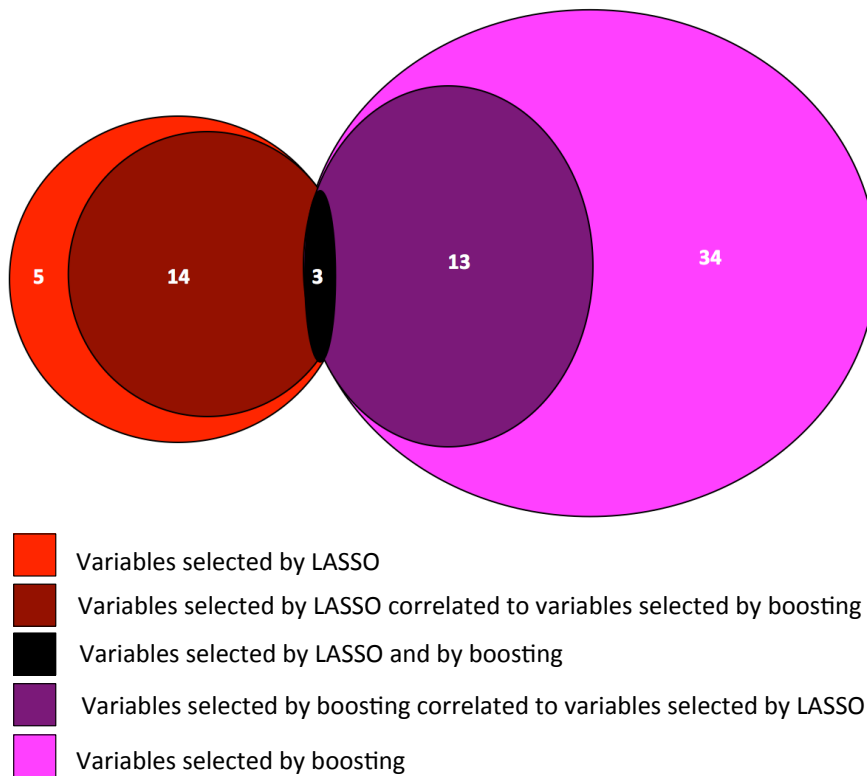


Figure A.13: Overlap of variables selected by LASSO and variables ranked among the top 50 by boosting with treating variables of the two algorithms as equal if their Pearson correlation coefficient was greater than or equal to 0.6. There were 30 variables fulfilling this criterion: 14 variables selected by LASSO and 13 variables from the top boosting variables, 3 by both.