Identification of biomarker-defined populations in precision medicine

Dissertation
zur Erlangung des humanwissenschaftlichen Doktorgrades
in der Medizin
der Georg-August-Universität Göttingen

vorgelegt von
Cynthia Laurena Huber
aus Stuttgart

Göttingen, 2023

Thesis Committee

First Supervisor:

Professor Dr. Tim Friede (Reviewer 1),
Institut für Medizinische Statistik, Universitätsmedizin Göttingen

Further Committee Members:

Prof. Dr. Jürgen Brockmöller,
Institut für Klinische Pharmakologie, Universitätsmedizin Göttingen

PD Dr. Norbert Benda,
Bundesinstitut für Arzneimittel und Medizinprodukte, Bonn

Further members of the examination board:

Prof. Dr. Heike Bickeböller (Reviewer 2),
Institut für Genetische Epidemiologie, Universitätsmedizin Göttingen

Prof. Dr. Tim Beißbarth,
Institut für Medizinische Bioinformatik, Universitätsmedizin Göttingen

Prof. Dr. Eva Hummers,
Institut für Allgemeinmedizin, Universitätsmedizin Göttingen

Prof. Dr. mult. Thomas Meyer,
Klinik für Psychosomatische Medizin und Psychotherapie,
Universitätsmedizin Göttingen

Date of the oral examination: 06.03.2023

Declaration of Authorship

I, Cynthia Laurena HUBER, declare that this dissertation titled, "Identification of biomarker-defined populations in precision medicine" and the work presented in it are my own and that it was written independently with no other sources and aids than quoted.

Abstract

The aim of precision medicine is to identify the treatment that provides the best response for a patient. For this purpose, predictive biomarkers play a crucial role. Due to their ability to define subgroups of patients that respond differently to treatment, they are highly useful. Biomarker-related subgrouping of the patient population may have different reasons, e.g. an improved benefit-risk balance, and can be supported by different sources of evidence, e.g. clinical data or pharmacological evidence investigating the biochemical or physiological effect of a drug on cells, organs, and systems. To assess the usefulness of a biomarker stratifying the patient population considering the presented evidence, a classification scheme with five increasing levels of evidence with regard to the expected molecular mechanism and the clinical evidence was proposed as part of this dissertation. Additionally, for each of the categories, an example of a biomarker-drug pair were suggested.

As the mechanism of action of a drug is not always fully understood or maybe even unknown, data-driven identification of differential treatment effects in subgroups suggesting treatment-by-subgroup or more precisely treatment-by-biomarker interactions is of interest to inform further research. Various data-driven subgroup identification methods have been proposed. However, neutral and systematic comparisons of their performance in simulation studies are rare. Therefore, I conducted a simulation study in order to compare five popular approaches regarding their capability to select a target population for subsequent trials. Although most of the methods performed well in settings with larger effects or more substantial sample sizes, all methods have difficulties in more realistic drug development settings with sample sizes that are not sufficiently large for identifying treatment heterogeneity across the population.

Pooling data from multiple trials can increase the sample size on which subgroup identification is performed. When pooling data from multiple studies, however, the between-trial heterogeneity must be taken into account, as otherwise spurious subgroups might be identified. Therefore, I proposed the metaMOB approach for subgroup identification in individual participant data (IPD) meta-analysis. The proposed approach combines commonly made assumptions in random-effects meta-analysis regarding between-trial heterogeneity and the generalized mixed-effects model tree algorithm based on model-based recursive partitioning (MOB). Using a Monte-Carlo simulation study, I showed that metaMOB is an appropriate subgroup identification method for IPD resulting from multiple heterogeneous trials.

Discrete time-to-event data needs specialised methods for data analysis including subgroup identification. Although MOB is applicable to a wide range of different outcome measures, e.g. normal, or binary, it is not suitable for discrete time-to-event data. For discrete time-to-event models which are based on binary outcome models, I could show that the type I error rate of the M-fluctuation test used as splitting criterion in MOB is inflated. I illustrated the inflated type I error rate in a simulation study and proposed a revised version of MOB for discrete time-to-event data, which is based on a resampling procedure and which controls the type I error more closely.

Acknowledgements

I would like to express my sincere gratitude to all those who helped me during my work on this dissertation. Without them, this work would not have been possible.
First and foremost, I would like to thank my supervisor Professor Tim Friede for his support throughout the development of this thesis. I am grateful for the helpful discussions, the suggestions to improve my drafts, both in terms of content and style, and for sharing his vast knowledge of clinical trials and beyond.
I would also like to thank Prof. Heike Bickeböller, who kindly agreed to be my second assessor. Furthermore, my sincere thanks are due to PD Dr. Norbert Benda for being a member of my thesis committee, for co-authoring my publications, and for supervising my internships at the Federal Institute for Drugs and Medical Devices (BfArM). In addition, I would like to express my gratitude to Prof. Jürgen Brockmöller for being a further member of my thesis committee. Moreover, I am also very grateful to Prof. Julia Stingl for co-authoring one of my papers and to Prof. Matthias Schmid for his valuable suggestions and ideas on my fourth paper.
Besides my advisors, I would like to thank all my (former) colleagues, and fellow PhD students, especially Burak, for the conducive working atmosphere on the top floor and the many cheerful moments in everyday working life.
For proofreading this thesis and providing helpful comments, I thank Tobias, Katharina and Steffen.
Last but not least, I want to thank everybody who supported me in my private life. In particular, I would like to thank my parents for their continuous support and unconditional love. Special thanks also go to my siblings and Markus for their support and understanding, and for the often-needed distraction.

Contents

Declaration of Authorship **iii**

Abstract **v**

Acknowledgements **vii**

1 Introduction **1**
1.1 Precision medicine and biomarkers 1
1.2 Research questions 4
1.2.1 Framework for classifying evidence for biomarker-driven patient selection 4
1.2.2 Comparison of subgroup identification methods 5
1.2.3 Subgroup identification in individual participant data meta-analysis 6
1.2.4 Subgroup identification for discrete survival data 7
1.3 Outline 8

2 Proposed approaches for identification of biomarker defined populations **9**
2.1 Classification framework for biomarker-drug pairs 9
2.2 Comparison of subgroup identification methods 11
2.3 Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning 19
2.4 Model-based recursive partitioning for discrete event times 23

3 Discussion **31**

Bibliography **35**

A Original Articles **43**
A.1 Classification of Companion Diagnostics: A new framework for biomarker-driven patient selection 43
A.2 A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations 43
A.3 Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning 43
A.4 Model-based recursive partitioning for discrete event times 43

List of Figures

1.1 Schematic plot for predictive and prognostic biomarker-defined subgroups. 2

2.1 Theoretical example of trees obtained by the different methods 15
2.2 Hypothetical example of the shape of three $BM+$ groups. 15
2.3 False discovery rate for model M0 as mean function of the data generating model. 17
2.4 Type II error rate vs sample size for three interaction effect sizes β_1 in setting with the step function model M1 using $\gamma = 0$. 18
2.5 False discovery rate for settings with $K = 5$, $\text{cor}(b_1, \mathbf{Z}) \neq 0$ and $\text{cor}(b_0, \mathbf{Z}) = 0$. 22
2.6 Fixed effect structure $f(\cdot)$ for the data generating model. Z_1,Z_2 and Z_5 denote the covariates defining the four subgroups. 23
2.7 Tree accuracy for model (M_3) as data-generating mechanism. 24
2.8 Illustration of the true survival functions used in the simulations for settings with $L = 6$. 27
2.9 Type I error rate for MOB and MOB-dS based on 2000 simulated data sets per setting with 20% censoring and 1000 permutation samples for MOB-ds. 28
2.10 Right-hand tail probabilities from approximate asymptotic and sampling distributions of the instability test in MOB (solid black) and MOB-dS (dashed red). 29

List of Tables

2.1 Parameters in the simulation study for assessing the selection accuracy. 16
2.2 Scenarios considered in the simulation study. 21

List of Abbreviations

ALS	Amyotrophic lateral sclerosis
ARDP	Adaptive refinement by directed peeling algorithm
$BM+$	Biomarker-positive
$BM-$	Biomarker-negative
CART	Classification and regression tree
e.g.	for example
EMA	European Medicines Agency
EGFR	Epidermal Growth Factor Receptor
FDA	Food and Drug Administration
FDR	False discovery rate
GLM	Generalised models
GLMM	Generalised mixed models
HER2	Human epidermal growth factor receptor 2
i.e.	that is
IPD	Individual participant data
IT	Interaction tree
KRAS	Kirsten rat sarcoma virus
Lasso	Least absolute shrinkage and selection operator
MOB	Model-based recursive partitioning
NSCLC	Non-small cell lung cancer
PRO-ACT	Pooled Resource Open Access ALS Clinical trials
SIDES	Subgroup identification based on differential effect search
STIMA	Simultaneous threshold interaction modelling algorithm

Chapter 1

Introduction

1.1 Precision medicine and biomarkers

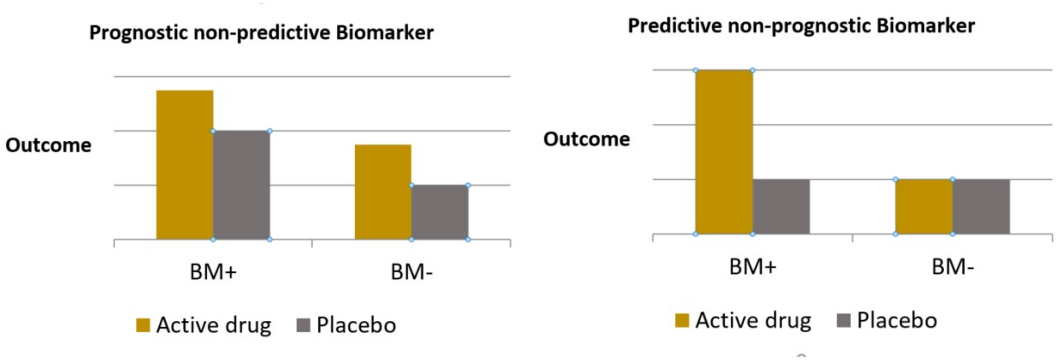
Precision medicine, often referred to as personalized medicine, aims at targeting the right treatments for the right patients at the right time [35, 74]. To treat a disease, the choice of drug or intervention should generally take into account both the disease and the characteristics of the patient to be treated. The treatment should promise the greatest benefit for the patient compared to alternative treatments and should have the least safety concerns. In drug development, commonly the *one size fits all* approach is used, meaning that the existence and magnitude of treatment and safety effects are established at a population-level, comparing outcomes in the same patients under different treatment conditions [46]. In contrast, precision medicine takes into account individual differences. Precision medicine does not mean that treatments are developed individually for each patient. Rather, the aim is to identify patients with similar treatment effects, which constitute subgroups. These subgroups are also expected to benefit particularly from the treatment compared to other subgroups. For instance, in retrospective analyses, the drugs panitumumab and cetuximab have been shown to be effective only in patients with the natural, non-mutated (unchanged) form, i.e. the wild-type, of the gene Kirsten rat sarcoma virus (KRAS) [61].

Precision medicine has become popular in recent years, as can be seen by the approvals of the Food and Drug Administration (FDA): The number of FDA approvals classified as personalised medicines by the Personalized Medicine Coalition [69] has increased from 28% in 2015 to 35% in 2021. Technological advances such as next-generation sequencing, which allow better characterisation of individuals through genetic and proteomic biomarkers, have contributed to the popularity alongside the US precision medicine initiative [87].

The term biomarker, however, does not necessarily refer to genes, proteins, or other biological molecules found in blood, other body fluids, or tissues. Although there are restrictive definitions of biomarkers, e.g. by the European Medicines Agency (EMA) [29], more general definitions are also common. The Biomarkers Definitions Working Group [9] defines a biomarker as a "characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention". In my dissertation, I use the term biomarker to refer to clinical covariates, e.g. demographics, disease severity scores, or other imaging, pharmacological, genomic, or proteomic biomarkers.

In drug development biomarkers can have different purposes. They can be used for the diagnosis of a disease, for predicting the course of a disease or the response to treatment, to stratify patient populations, to monitor patients, or as an endpoint

in clinical trials [30]. For precision medicine, biomarkers that are able to discriminate individuals with different treatment or safety profiles are of particular interest. These biomarkers are called predictive biomarkers and have to be distinguished from prognostic biomarkers which predict the natural course of a disease, e.g. the likelihood of a clinical event. A prognostic biomarker can discriminate individuals with different baseline risks for a clinical event of interest [30]. Figure 1.1 illustrates the difference of prognostic and predictive biomarkers with a hypothetical example. The outcomes of the experimental and control groups are shown for two disjoint biomarker-defined subgroups. The two subgroups are referred to as biomarker positive (BM+) and biomarker negative (BM-). In the following, BM+ refers to the subgroup in which the outcome is better or the treatment success is greater. For subgroups defined by prognostic biomarkers, the outcome is different in the two biomarker groups irrespective of the treatment group. The difference between the treatment groups in BM+ and BM- is the same (see Figure 1.1 (A)). For purely predictive biomarkers, however, a treatment difference between the BM+ and BM- groups is observed (see Figure 1.1 (B)). Therefore, predictive biomarkers are drug specific, whereas prognostic biomarkers are not.



(A) Prognostic (B) Predictive

FIGURE 1.1: Schematic plot of the outcome for the experimental and control treatment in biomarker-defined subgroups. Larger outcome values are better. (A) The biomarker is prognostic only and (B) the biomarker is predictive only. BM+, biomarker positive subgroup; BM-, biomarker negative subgroup.

Although a *one size fits all* approach is often employed in the development of medicines, it is acknowledged that the treatment might be heterogeneous across subgroups. The regulatory agencies FDA and EMA issued guidelines [24, 33] related to subgroup analyses. The EMA guideline on subgroup analysis focuses on exploratory subgroup analyses [24]. Therefore, guidance on confirmatory subgroup analysis involving one or two prospectively defined subgroups for evaluating the efficacy profile of a new treatment can be found in other documents, e.g. in the guideline on multiplicity issues [22] or in the *Reflection paper on methodological issues in clinical trials planned with an adaptive design* [21]. For confirmatory subgroup analyses the type I error rate is commonly controlled and only one or two prespecified subgroups are considered [57].
Exploratory subgroup evaluation and post-hoc subgroup evaluation as defined by Lipkovich et al. [57] consider a relatively small number of prespecified subgroups. The EMA guideline distinguishes three scenarios in which exploratory subgroup analysis might be pursued: (1) clinical data are overall statistically persuasive with

the primary analysis showing therapeutic efficacy, (2) clinical data are statistically persuasive in the primary analysis, but the efficacy or risk-benefit is borderline or unconvincing, and (3) clinical data failed to establish treatment efficacy in the primary analysis.
In confirmatory trials that showed statistically persuasive results in the primary analysis population, the consistency of efficacy or safety is routinely assessed. In fact, health authorities require these exploratory subgroup analyses to investigate small numbers of predefined subgroups of the overall patient population, e.g. gender, age, region, and known prognostic biomarkers [24, 47]. The analyses of treatment heterogeneity form the basis for assessing the consistency and the applicability of the treatment effect to the patient population. Label restrictions may even apply to an inconsistent treatment effect or an unacceptable safety profile in a subgroup. In the case of Gefitinib (brand name IRESSA), the initial market authorisation application at the EMA was for the treatment of adults with locally advanced or metastatic non-small cell lung cancer (NSCLC). Based on the results of a post-hoc analysis of a study for Gefitinib showing an interaction between Epidermal growth factor receptor (EGFR) tyrosine kinase and treatment the indication was restricted to adult patients with locally advanced or metastatic NSCLC with activating EGFR tyrosine kinase [27].
In failed confirmatory trials or trials with borderline treatment efficacy or risk-benefit as described by the subgroup guideline of the EMA [24], subgroup analysis and the identification of subgroups with increased or decreased treatment effect are also of interest. However, confirmatory conclusions are no longer possible if the primary null hypothesis cannot be rejected and further studies are needed [24]. Additionally, exploratory subgroup selection based on overall non-significant trials should be performed with caution because the subgroup treatment effect estimate is usually associated with a high bias [85].
The exploratory subgroup analysis plan defined in the study protocol should include prespecified subgroups defined by biomarkers, which are assumed to be prognostic or which are assumed to modify the treatment effect [47]. The inclusion of subgroups in the study protocol requires prior hypotheses regarding a biomarker being prognostic or predictive. The mechanism of action of a drug generates such hypotheses regarding treatment heterogeneity. Especially in oncology, modern drugs are developed in order to act on specific genetic targets. Therefore, it can be expected expected that the drug is only effective in individuals in whom the target is present. The genetic biomarker, the target, can be considered predictive because it causes differences in treatment efficacy between patients with and without the biomarker, e.g. a gene mutation. If the drug's mechanism of action is not fully understood, statistical approaches for identifying subgroups with differential treatment effect are of interest in order to generate hypotheses. Thus, they help to inform the design of future trials, e.g. multi-population designs or adaptive enrichment designs [7, 14, 33, 36, 49, 90]. The aim of subgroup identification or subgroup discovery, as referred to by Lipkovich et al. [57], is to select subpopulations in which treatment efficacy or safety is improved. Commonly, methods for identifying subgroups use a larger number of biomarkers to form subgroups with a differential treatment effect. Statistically speaking, identifying subgroups with different treatment effects is equivalent to identifying treatment-by-subgroup interactions. Subgroups are usually defined by a single biomarker or a combination of biomarkers. Biomarkers that are involved in the interaction and thus define the subgroup are predictive. However, the identification of subgroups can be difficult in practice. This is due to the fact that clinical trials are usually not powered to detect interactions. Moreover, to define subgroups

cut-off values for continuous biomarkers have to be chosen. When using interaction tests, different cut-off values increase the number of subgroups to be tested and thus the risk of chance findings. Moreover, additional cut-off selection increases the risk of misclassifying patients into either the BM+ or BM- group. Another challenge arising with subgroup identification is the overly optimistic estimated treatment effect within the identified subgroups. This is particularly relevant when planning subsequent studies which are based on estimated treatment effects in the identified patient populations. If these estimates are biased, this can lead to wrong decisions in study planning, resulting in wasted time and money or missed opportunities for both sponsors and society [57]. In addition, subgroups with lower prevalence may not be very attractive for further planning by the sponsor.

1.2 Research questions

In my research, I focused on the identification of subgroups differing with respect to their treatment benefit. In short, I first examined the types of evidence that support the plausibility and utility of biomarkers included in the subgroup definition when it comes to their ability to discriminate individuals with differential treatment effects. Second, I focused on statistical methods for the identification of subgroups in precision medicine. To that end, I compared five machine learning approaches in a simulation study. Third, I proposed a subgroup identification approach for individual participant data from multiple studies and an approach tailored to discrete time-to-event event outcomes. In the following, I outline my motivation for investigating these issues in detail.

1.2.1 Framework for classifying evidence for biomarker-driven patient selection

Biomarker-based patient selection in clinical trials can have various reasons. The biomarker may improve the disease definition or prognosis, or it may suggest a better treatment effect or a reduction of adverse effects [23]. Although the benefit of this restriction does not have to be demonstrated for the approval of drugs, regulatory considerations relate to the usefulness of this restriction. Often, however, the evidence for justifying the patient selection is sparse or unclear. Moreover, the sources of this evidence are diverse and its relevance may be vague.
It is common practice to stratify patients into biomarker-defined subgroups due to improvements in the safe and effective use of drugs based on prior knowledge of their mechanism of action. In the absence of a full understanding of the mechanism of action, data-driven evidence of treatment-by-subgroup interactions may be used to support the pharmacological and biological reasoning of the predictive nature of a biomarker. Data-driven evidence for treatment-by-subgroup interactions is scarce, because clinical trials are usually not powered to detect statistically significant interactions. Additionally, for drug approval, it is not necessary to demonstrate the benefits of this restriction. The efficacy and tolerability of a drug has to be demonstrated only within the intended patient population.
Nevertheless, in the case of a treatment approved in a biomarker-positive subgroup, extrapolation of the benefit-risk ratio to the biomarker-negative population may be considered at a later stage to expand the indicated patient population. To make such an extrapolation, however, data from biomarker-negative patients are necessary, and the biological plausibility of the benefit in this subgroup must also be considered. For instance, the mode of action in different biomarker-drug combinations may be

different, so data on treatment in biomarker-negative patients may not be necessary or even ethical. Therefore, the benefit of a biomarker for precision medicine is defined by biological plausibility combined with empirical evidence from clinical and non-clinical studies [24].
To better assess the usefulness of biomarker-based subgrouping, a classification of the level of evidence is needed that distinguishes between empirical and biological evidence for patient selection. My work [44] proposed such a framework allowing for a classification of the underlying evidence that can support regulatory and scientific decision-making with respect to biomarker-based selections. In addition, for the proposed categories, drugs approved by the EMA with biomarker information on the label were classified into the proposed categories based on the evidence provided in their European Assessment Reports or the Summary of Product Characteristics.

1.2.2 Comparison of subgroup identification methods

A variety of methods have been presented in the past few years for the data-based identification of subgroups with differential treatment effects. The systematic review by Ondra et al. [66] found 86 articles on the identification and confirmation of targeted subgroups. The tutorial by Lipkovich et al. [57] includes around 60 methods for subgroup identification. For the identification of treatment-by-subgroup interactions, tree-based methods are popular, as they identify predictive biomarkers and select a cut-off value for continuous biomarkers to define decision rules. Interaction trees (IT) [83], model-based recursive partitioning (MOB) [77], subgroup identification based on differential effect search (SIDES) [58], and the simultaneous threshold interaction modelling algorithm (STIMA) [20] are examples of tree-based methods already described elsewhere that aim at identifying subgroups with an increased treatment effect. The adaptive refinement by directed peeling algorithm (ARDP) for subgroup identification, as included in Patel et al. [68], is also an example of a subgroup identification method identifying predictive biomarkers with their corresponding cut-off value. Despite the availability of a wealth of methods, there are few neutral comparative studies focusing on the comparison itself. Boulesteix et al. [11, 12, 13] stress the need for neutral comparisons that compare existing methods described elsewhere and are conducted by a group of researchers that are (ideally) equally familiar with the methods being compared. Studies published before Huber et al. [42], i.e. [1, 19, 79], focused on emphasizing the differences between methods and did not always take into account scenarios that are relevant in drug development. In drug development, it is often of interest to identify one subgroup with a larger treatment effect, e.g. to inform the trial design of future studies in order to increase the probability of a successful subsequent trial. The results of the subgroup identification methods that I included in the comparison study are used to determine which subgroups will comprise the target population, that is, a subgroup with a larger treatment benefit. I proposed a subgroup criterion for this selection [42]. The treatment effects in each of the identified subgroups have to exceed a prespecified threshold in order to be assigned to a potential future target population, the biomarker-positive (BM+) subgroup. Using the proposed subgroup criterion I compared five subgroup identification methods in my work [42], namely, IT, MOB, SIDES, STIMA, and ARDP in the situation of a randomized controlled clinical trial. By using the subgroup criterion, comparable results are obtained for these methods.

MOB, IT, and SIDES are able to identify multiple subgroups, ARDP identifies a sequence of potential subgroups, and the application of STIMA results in a regression model with interaction terms. To evaluate which method selects the $BM+$ subgroup best in which situation, the methods were compared in a Monte Carlo simulation study.

1.2.3 Subgroup identification in individual participant data meta-analysis

Comparison studies on subgroup identification methods including my work [42] showed that sample size is one of the important factors influencing the performance of subgroup identification methods as measured by different criteria, including correctly identifying treatment-by-biomarker interactions, and false discovery rate, i.e. spurious identification of a subgroup although the treatment effect is homogeneous across the entire population [1, 42, 79]. Clinical trials are usually not powered to detect predefined treatment-by-subgroup interactions in exploratory subgroup analyses [16], thus it is not surprising that data from one clinical trial are often also not sufficient for the identification of subgroups with differential treatment effects. This might not only apply to clinical trials, but also to other experiments such as studies in social sciences or preclinical experiments in life sciences. Pooling data from multiple trials may increase the sample size. For instance, repositories of individual participant data (IPD) with the aim of identifying subgroups of patients with differential treatment effects were developed by Patel et al. [67] for low back pain and the International Weight Management in Pregnancy Collaborative group [86] for gestational weight gain. The investigation of differential effects in subgroups of active scheduling as behavioural treatment for depression based on sixteen studies was conducted by Cuijpers et al. [18].
When pooling data from multiple studies in an IPD meta-analysis, it is important to take into account heterogeneity between studies resulting from, for example, differences in study designs, study populations, study quality, choice of comparator intervention, or other study-specific influences. Two (very different) types of between-trial heterogeneity may be present in data: heterogeneity in the baseline (control group) outcomes as well as heterogeneity in treatment effect sizes. Although Jackson et al. [48] consider only aggregate meta-analysis models and binary outcomes in their review article on random-effects models for meta-analyses, the approach to account for heterogeneity between studies in terms of treatment effect is the same in all models considered: Treatment effects are assumed to be normally distributed. The models in Jackson et al. [48] differ in their assumptions regarding the heterogeneity of baseline risks or the independence of treatment and baseline random effects. To account for heterogeneity in the baseline, the intercept term can be either modelled by separate intercepts for each trial, so-called stratified intercepts, or by random intercepts. While stratified intercepts do not require assumptions about the distribution of intercepts across studies, a (parametric) random intercept approach requires fewer parameters to be estimated [72]. Generalised mixed models (GLMM) are increasingly used for IPD meta-analyses [80], as they do not require aggregation of IPD data to be analysed in a second step with known and appropriate meta-analysis models, e.g. Jackson et al. [48].
GLMMs are the basis for most of the proposed methods of subgroup identification on IPD data accounting for trial heterogeneity [51]. However, the approaches published so far are not sufficiently flexible, e.g. Wang et al. [91, 92], and also tend to model heterogeneity between studies using simpler models, e.g. Mistry et al. [62]

and Fokkema et al. [32]. The approaches by Wang et al. [91, 92] investigate the treatment heterogeneity across one continuous biomarker only by either a fixed-effect [91] or a random effect [92] meta-analysis approach. The detection of interactions of multiple biomarkers with the treatment indicator variable is not possible using this framework. Tree-based subgroup identification methods accounting for between-study heterogeneity allow the detection of more complex interactions. However, the tree-based methods by Mistry et al. [62] extending SIDES [58] and by Fokkema et al. [32] extending MOB [96, 77] only consider heterogeneity in the baseline. Ignoring between-trial heterogeneity can lead to the identification of wrong splitting variables and thus wrong definitions of the subgroups [78].
The aim of my research was to introduce and assess an approach that systematically combines commonly made assumptions in meta-analysis models with the GLMM-tree algorithm by Fokkema et al. [32]. GLMM-trees are based on MOB [77, 96] which showed a good or even the best performance in various neutral comparison studies [1, 42, 59, 79]. In particular, I focused on using different assumptions for the between-trial heterogeneity in the baseline in the introduced metaMOB approach and their impact on the performance regarding the false discovery rate and correctly identifying subgroups.

1.2.4 Subgroup identification for discrete survival data

Event times are sometimes not measured on a continuous but a discrete scale. In some situations, it might be measured on a rather coarse scale (e.g. grouped or rounded event times). Common examples of discrete event times are clinical or epidemiological studies with a fixed number of follow-up visits not allowing for continuous monitoring of event times, e.g. [81]. The time to infection, e.g. [6], or time to death, e.g. [39], of patients collected on a daily basis in intensive care units are other examples of event times measured on a discrete scale. Intrinsically discrete time duration is encountered in studies investigating the time to pregnancy assessed by the number of menstrual cycles, e.g. [73, 31].
The discreteness of the event times should be accounted for appropriately as, otherwise, the estimators could be biased and the predictions inaccurate. Furthermore, methods for continuous event times assume that there are no “ties”. Ties are observations with the same event time, which commonly occur in discrete time-to-event data. Methods appropriate for the analysis of discrete event times have been proposed by Tutz and Schmid [89], Willet and Singer [94] and Schmid and Berger [75]. Tree-based models including random forest approaches specifically for discrete time-to-event data have been proposed by Schmid et al. [76], Bou-Hamad et al. [10] and Moradian et al. [63]. These methods aim to identify subgroups defined by prognostic markers by partitioning the data based on impurity measures instead of a formal statistical test. The semi-parametric model-based recursive partitioning approach, however, relies on formal statistical hypothesis testing to identify the subgroups defined by both prognostic and predictive biomarkers. Additionally, it is applicable to a broad range of outcome measures. Besides continuous outcomes as assumed in two of my publications [42, 43], MOB can also be used for binary or count data, as the splitting criterion and the node-wise fitting in MOB is based on generalised linear models. MOB is also applicable to survival regression [96], but it considers the time-to-event outcome to be continuous throughout. As the likelihood of a discrete time-to-event model is equivalent to the likelihood of a binary regression model, it seems that MOB can be readily applied to discrete time-to-event data by using the standard implementation for MOB [41]. MOB, however, controls for the

percentage of incorrectly identified subgroups in binary models with independent subjects. The percentage of falsely identifying subgroups although none are present is controlled by the test used as splitting criterion in MOB, namely the M-fluctuation test, and by the use of multiplicity adjustments [42, 79, 59]. The M-fluctuation test investigates parameter instabilities of the regression model assumed for MOB [95, 96]. As discrete survival models are fitted using an augmented data matrix, the rows of the matrix are not independent as subjects are commonly represented by multiple rows. Therefore, the assumption of independent observations made for the M-fluctuation test is violated and the asymptotic theory of the test is not valid for discrete event times.
In my research I showed that applying the standard MOB (for binary data) to discrete time-to-event data, i.e. ignoring dependencies in the augmented data matrix, leads to a systematic inflation of the type I error rate. Therefore, I developed a method for identifying subgroups defined by either prognostic, predictive, or prognostic and predictive biomarkers specifically tailored to modelling discrete survival data that controls the type I error rate better compared to the original MOB procedure. The method employs a permutation procedure that accounts for the structure of the augmented data matrix used for modelling discrete time-to-event data.

1.3 Outline

In this dissertation, I address the research questions described in Chapter 1.2. The results of my research have either been published in peer-reviewed journals [42, 43, 44] or are currently under review [45]:

- [42] Huber, C., Benda, N., and Friede, T. "A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations". In: *Pharmaceutical Statistics* 18.5 (2019), pp. 600–626
- [43] Huber, C., Benda, N., and Friede, T. "Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning". In: *Advances in Data Analysis and Classification* 16.3 (2022), pp. 797–815
- [44] Huber, C., Friede, T., Stingl, J., and Benda, N. "Classification of Companion Diagnostics: A New Framework for Biomarker-Driven Patient Selection". In: *Therapeutic Innovation & Regulatory Science* 56.2 (2022), pp. 244–254
- [45] Huber, C., Schmid, M., and Friede, T. *Model-based recursive partitioning for discrete event times*. (2022). URL: `http://arxiv.org/pdf/2209.06592v1`

In Chapter 2, I present the details of my work addressing the research questions outlined in Chapter 1.2. Chapter 3 includes a critical discussion of the chosen approaches and proposed statistical methods. Furthermore, I provide an outlook on future research regarding subgroup identification for precision medicine.

Chapter 2

Proposed approaches for identification of biomarker defined populations

2.1 Classification framework for biomarker-drug pairs

The results of my research on different types of evidence for the plausibility and usefulness of a biomarker-based patient selection are published in Huber et al. [44]. Patient selection is commonly based on a companion diagnostic, that is a medical device, which is often an in vitro device. This diagnostic device identifies patients that are suitable or unsuitable for a safe and effective use of a specific drug. In my work I propose a classification scheme for the evidence of a biomarker's predictive value in relation to a specific drug. The proposed classification considers two dimensions for biomarker-drug pairs. First, the pharmacological mechanism and the biological plausibility of the biomarker being the drug target, and second the evidence with respect to clinical data. Furthermore, a distinction is made between comparative and non-comparative evidence for these two dimensions. Comparative evidence can be obtained from animal or clinical studies in which the treatment is compared to a control or in which the biomarker is associated with the drug's mechanism of action, e.g. the biomarker characterizes the drug target. Therefore, a differential effect in biomarker-positive and biomarker-negative patients can be assumed. Comparative clinical evidence allows us to evaluate the interaction between drug and biomarker. The proposed classification consists of five categories with increasing evidence for a biomarker-based patient selection:

A Biomarkers with non-comparative evidence:
Pharmacological mechanisms indicate a prognostic value of the biomarker which is unrelated to the drug target

B Biomarkers with comparative clinical data only:
Comparative clinical data are available indicating a prognostic value

C Biomarkers with comparative pharmacological evidence:
Pharmacological mechanism indicates treatment benefit depending on biomarker expression

D Biomarkers with comparative pharmacological and clinical evidence:
Pharmacological evidence and clinical data suggest a differential treatment effect in biomarker-defined subgroups

E Biomarkers demonstrated to be predictive in confirmatory clinical studies:
Evidence generated by confirmatory study stratifying patients by biomarker

status to demonstrate significant and relevant interaction between treatment and biomarker-defined subgroup.

Although comparative evidence is available for biomarkers of the biomarker-drug pairs categorized in *B*, the level of evidence for the biomarker being predictive is weak. Larger studies with stratification by biomarker status in order to demonstrate the predictive effect of the biomarker are needed. However, conducting such studies does not seem realistic without a biological rationale for a differential treatment effect in the biomarker-defined subgroups. Biomarker-drug pairs of category *C* are reasonable candidates for further investigation of the predictive value of the biomarker in subsequent trials. For some biomarker-drug pairs, though, the generation of comparative clinical data and an increase in evidence level may not be necessary or even possible due to ethical concerns if pharmacological evidence suggests adverse effects on biomarker-negative patients.
For biomarker-drug pairs of category *C*, *D*, and *E* aspects such as the pathogenic mechanism, the relationship of biomarker and efficacy, and biomarker and safety have to be considered. My work distinguished between three cases relating to categories *C*, *D*, and *E* [44]:

- Detection of the specific pathological change implying zero efficacy if the biomarker is negative, e.g. BCR-abl for the drug Imatinib [26].
- Measures of the activation of a pathway (yes/no), implying that the drug is more likely to be efficacious if the biomarker is positive. However, efficacy ultimately depends on other surrounding factors, e.g. tumour heterogeneity. The biomarker might therefore not be valid in all contexts, e.g. KRAS is valid as a biomarker in colon cancer but not in lung cancer for detecting an inactive epidermal growth factor receptor (EGFR) pathway in treatment with EGFR inhibitors [53].
- Indirect measure of overexpression implying a higher efficacy if the biomarker is positive. However, biomarker-positive non-tumour cells may have safety implications, e.g. HER2 for the drug trastuzumab [25].

For the proposed categories I identified ten biomarker-drug pairs approved in the EU based on the PharmGKB Database [93], which collects knowledge about the impact of genetic biomarkers on drug responses and identifies drugs with labels containing pharmacogenetic information. For the categorization of the evidence information published in the documents such as the European Assessment Reports or Summary of Product Characteristics which are available on `https://www.ema.europa.eu/en` are used.
Only for category *E* describing the highest level of (confirmatory) evidence, no biomarker-drug pair could be identified. However, the categorization of biomarker-drug pairs can change over time, if further studies on the drug and biomarker are carried out. An example of category *C* is the drug osimertinib together with the biomarker EGFR. Osimertinib was approved by the EMA and FDA for advanced non-small cell lung cancer (NSCLC) as first-line treatment for patients with activating EGFR mutations or patients with locally advanced or metastatic EGFR T790M mutation-positive NSCLC [28, 34]. The inhibitory activity against EGFR was demonstrated in vitro and tumor shrinkage was shown in mouse lung tumor models [28]. Clinical studies of osimertinib did not include patients without EGFR mutations. Therefore, no comparative clinical evidence is available which would justify a classification into a higher category. Further examples of biomarker-drug pairs and the

evidence category they belong to can be found in Huber et al. [44].
In summary, I outlined a classification scheme for biomarker-drug pairs indicating increasing evidence of the usefulness of a biomarker stratifying the patient population. The categories are based on both pharmacological and data-based evidence. The classification is useful for deciding whether the presented evidence is appropriate for justifying biomarker-defined patient selections. For each proposed category except category *E*, examples of biomarker-drug pairs were identified. To conclude, the classification scheme can help to strengthen and focus discussions in regulatory authorities on the qualification of new biomarkers and improve the comparability of different biomarker-drug pairs.

2.2 Comparison of subgroup identification methods

Although numerous approaches for the identification of subgroups with differential treatment effects are available [57], comparisons of these approaches are still scarce. In the following, I summarize the main results of the comparison of five subgroup identification methods published in Huber et al. [42]. First, I outline the methods. Second, I introduce the criteria used for assessing the performance of the methods. Last, I summarize the key findings of the simulation study.
For the comparison study methods were selected that are applicable to a continuous outcome and also include a cut-off selection for continuous biomarkers. The five methods included in the comparison study are Interaction tree (IT) [83], model-based recursive partitioning (MOB) [77], subgroup identification based on differential effect search (SIDES) [58], simultaneous threshold interaction modelling algorithm (STIMA) [20], and adaptive refinement by directed peeling algorithm (ARDP) [68].

Methods

In the comparison study only the situation of a randomized controlled clinical trial is considered. The data consist of N independent and identically distributed observations $\{(y_i, a_i, \mathbf{z}_i) : i = 1, \ldots, N\}$. The data includes the continuous outcome variable Y, the treatment indicator A, with $A = 1$ for the experimental and $A = 0$ for the control group, and p covariates $\mathbf{Z} = (Z_1, \ldots, Z_p)$. The observed values for subject i of Y, A and $\mathbf{Z}$ are denoted by y_i, a_i and $\mathbf{z}_i$, respectively. It is assumed that the covariates $\mathbf{Z}$ potentially define subgroups with differential treatment effects. Therefore, the covariates $\mathbf{Z}$ are investigated with regard to a potential interaction with the treatment effect. Without loss of generality, it is assumed that larger values of the outcome Y are preferable. The expected outcome is denoted by μ.

Interaction trees start with growing a large initial tree following the Classification and regression trees (CART) method [15] using the squared t-test for testing the null hypothesis $H_0 : \beta_1 = 0$ of the following linear regression model as splitting criterion:

$$\mu_{ij} = \alpha + \beta_0 \cdot a_i + \gamma \cdot I(z_{ij} \leq c) + \beta_1 \cdot a_i \cdot I(z_{ij} \leq c) \text{ for } i = 1, \ldots, N \text{ and } j = 1, \ldots, p,$$

with μ_{ij} denoting the expected outcome of subject i for the model considering covariate Z_j. The indicator function is denoted by $I(\cdot)$. The biomarker Z_j with its corresponding cut-off value c yielding the maximum t^2 test statistic value is selected

for partitioning the data. The interaction-complexity criterion by Su et al. [83] is applied for pruning the initial interaction tree and for selecting the best-sized subtree from the resulting sequence of nested subtrees. In order to reduce the overoptimism in the results induced by applying the same complexity criterion to both pruning and final tree selection, an independent subset of the data is used for selecting the final tree.

Model-based recursive partitioning seeks to improve the model fit by partitioning the data with respect to some biomarkers $\mathbf{Z}$ and fitting separate local regression models of the form

$$\mu_i = \alpha + \beta_0 \cdot a_i \tag{2.1}$$

within each of the resulting partitions, where the expected outcome of subject i is denoted by μ_i. The data are partitioned if the model parameters α and β_0 differ across resulting partitions. This is assessed via the M-fluctuation test [95]. MOB applies a pre-pruning procedure, i.e. the data are only partitioned if the at least one of the null hypotheses of parameter stability of α or β_0 across $Z_j (j = 1, \ldots, p)$ can be rejected. After the selection of the partitioning variable Z_{j*}, the cut-off value is selected to maximize the sum of the log-likelihoods of the models in the two resulting subsets.

Simultaneous threshold interaction modelling algorithm uses a linear regression model for modelling the main effects and a tree for modelling higher-order interactions. For combining theses two approaches the tree is embedded in a reference model. For subgroup identification the first split is forced on the treatment indicator variable. Therefore, the initial reference model of STIMA which includes the tree in the model equation is defined as follows (Equation (2.2)):

$$\mu_i = \alpha + \beta_0 I(a_i = 1) + \sum_{j=1}^{p} \gamma_j z_{ij}. \tag{2.2}$$

The splitting procedure compares the current reference model, e.g. Equation (2.2), to an expanded model based on a tree with an additional split, e.g. Equation (2.3).

$$\mu_i = \alpha + \beta_0 I(a_i = 0) + \beta_1 I(a_i = 1) I(z_{ij*} > c^*) + \sum_{j=1}^{p} \gamma_j z_{ij}. \tag{2.3}$$

The variable Z_{j*} and cut-off value c^* defining the splits of the tree are selected by evaluating the relative increase in variance accounted by an expanded model. The splitting procedure is recursively applied using the model with the selected split as new reference model until a predefined stopping criterion is met. STIMA applies the pruning procedure for CART [15] on the obtained tree.

Subgroup identification based on differential effect search evaluates a splitting criterion in order to identify M best splits ($M = 3$ in the simulation study) of each node. The splitting criterion is defined as

$$p_{\text{SIDES}} = 2 \min(1 - \Phi(T_{\text{left}}), 1 - \Phi(T_{\text{right}})), \tag{2.4}$$

with T_{left} and T_{right} denoting the test statistics for a one-sided test of no differential treatment effect in the resulting left and right child nodes, respectively. The cumulative distribution function of the standard normal distribution is denoted by $\Phi(\cdot)$. The splitting procedure only considers covariates that do not define the parent nodes. Furthermore, only the child node of the identified pair of nodes with the larger treatment benefit is retained, provided the p-value is significant at a one-sided nominal level which is found using a resampling-based method. The splitting is repeated in the retained child nodes until no further predefined improvement of the test statistic value in the resulting child nodes compared to the parent nodes can be observed or a maximum tree depth or a prespecified minimum node size is reached.

Adaptive refinement by directed peeling does not partition the data but peels off observations in each iteration step resulting in a sequence of nested subgroups which can be illustrated by a tree. LeBlanc et al. [54] originally introduced ARDP for detecting subgroups with poor prognosis. An adaption for treatment-by-biomarker interactions was introduced in Patel et al. [68].
In iteration step r for each variable Z_j, a prespecified number of observations is peeled off in the direction indicated by the model

$$\mu_i = \alpha + \beta_0 a_i + \sum_{j=1}^{p} \beta_j z_{ij} a_i + \gamma_j z_j.$$

To increase the estimated treatment effect within the resulting region S_j^r, e.g. for a positive sign of the interaction $\beta_j (j = 1, \ldots, p)$, smaller values of Z_j are peeled off as larger values of the covariate Z_j lead to larger treatment effects.
The region S_{j*}^r with the largest treatment effect compared to the subgroup selected in the previous iteration step S^{r-1} is selected as the subgroup of the current iteration S^r. This procedure is repeated until a prespecified minimum size of the identified region is reached. In contrast to the other methods, the ARDP algorithm does not provide a pruning or selection criterion for the final subgroup. Instead a sequence of nested subgroups of subjects benefiting from the experimental treatment is obtained. In order to choose one of those subgroups as the final subgroup a selection criterion is needed which is described in the following.

Subgroup criterion

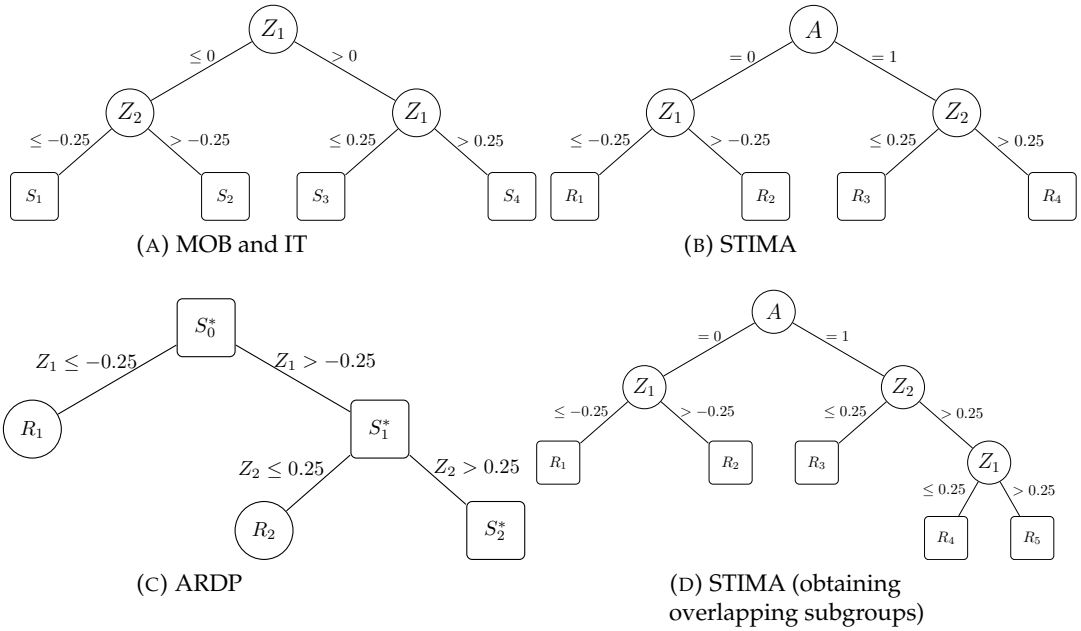
In order to compare the different methods, I describe how to derive a subgroup definition for different methods and how to define a potential target population. The target population is denoted by $BM+$, the biomarker-positive subgroup, which is assumed to benefit more from the experimental subgroup than its complement, the biomarker-negative subgroup, $BM-$.
MOB and IT identify disjoint subgroups whose definition can directly be derived from the splits of the estimated tree. In Figure 2.1(A) an example for MOB and IT with four subgroups denoted S_1, S_2, S_3 and S_4 is illustrated. The first subgroup S_1 is defined by the splits on $Z_1 \leq 0$ and $Z_2 \leq 0.25$ and therefore contains only subjects fulfilling these conditions. As STIMA forces its first split to be on the treatment indicator variable A in order to identify subgroups with heterogeneous treatment effects, the terminal nodes denoted by $R_1, \ldots, R_4$ in Figure 2.1(B) or by $R_1, \ldots, R_5$ in Figure 2.1(D) include only subjects assigned to either the experimental or control treatment. In my work [42] each terminal node of the experimental treatment

branch is combined with each terminal node of the control treatment branch to obtain subgroups. An example on how to derive subgroup definitions from the output of STIMA is given in Huber et al. [42]. STIMA can also yield overlapping subgroups, see Figure 2.1(D): Combining region R_1 with R_3 and R_1 with R_4 leads to nested subgroups. The interpretation of the subgroups identified with SIDES is equivalent to the interpretation of terminal nodes identified by MOB or IT. However, SIDES grows multiple trees in each iteration step resulting in the possibility of obtaining overlapping subgroups.

ARDP results in a sequence of potential subgroups. However, it does not provide "final" subgroups as the other methods do. In the hypothetical example presented in Figure 2.1(C) three potential subgroups denoted by S_0^*, S_1^*, and S_2^* and their corresponding complementary subgroups denoted by R_1 and R_2 are shown.

None of the methods described above identifies by default two complementary subgroups, $BM+$ and $BM-$. To obtain a potential future target population, $BM+$, the results of the subgroup identification methods need to be dichotomized. Since the $BM+$ subgroup should benefit from the experimental subgroup, it should only consist of subgroups whose estimated treatment effects exceed a prespecified threshold denoted by Δ_{mintrt}. In each identified subgroup $\hat{S}$ the treatment effect is estimated by the difference $\Delta(\hat{S}) = \mu_1(\hat{S}) - \mu_0(\hat{S})$, with $\mu_0(\hat{S}) = E(Y|A = 0, \mathbf{Z} \in \hat{S})$ and $\mu_1(\hat{S}) = E(Y|A = 1, \mathbf{Z} \in \hat{S})$ denoting the expected outcomes in the identified subgroups for subjects in the experimental or control arm. All subgroups identified by either IT, MOB, STIMA or SIDES meeting the criterion $\hat{\Delta}(\hat{S}) > \Delta_{\text{mintrt}}$ are combined into the $BM+$ group. Subjects not included in the $BM+$ group form the $BM-$ group. For ARDP, the largest subgroup out of the identified sequence of potential subgroups meeting the subgroup criterion is chosen as $BM+$.

Selecting a $BM+$ and $BM-$ subgroup based on the proposed subgroup criterion can lead to different shapes of the formed subgroups. A $BM+$ definition resulting from merging subgroups identified by MOB, IT, or STIMA can lead to a disjointed definition of the $BM+$ subgroup, illustrated by subgroups a and c, respectively, of the hypothetical example in Figure 2.2. In principle, subgroup b can be obtained by all methods, but this box shape of the $BM+$ is guaranteed only for ARDP. For the other methods, a simple tree, which can be obtained by an appropriate choice of certain tuning parameters, could also lead to a complex form of the $BM+$ group, as illustrated by subgroups a and b, respectively, in Figure 2.2. However, SIDES only allows one cut-off value for a single covariate in the definition of the subgroup, meaning that two cut-off values are not possible for one covariate. Therefore, SIDES results in half-open box shapes. These half-open boxes, which form the $BM+$ group, can also be disjoint, similar to subgroups a and c of Figure 2.2.



(A) MOB and IT

(B) STIMA

(C) ARDP

(D) STIMA (obtaining overlapping subgroups)

FIGURE 2.1: Theoretical example of trees obtained by the different methods

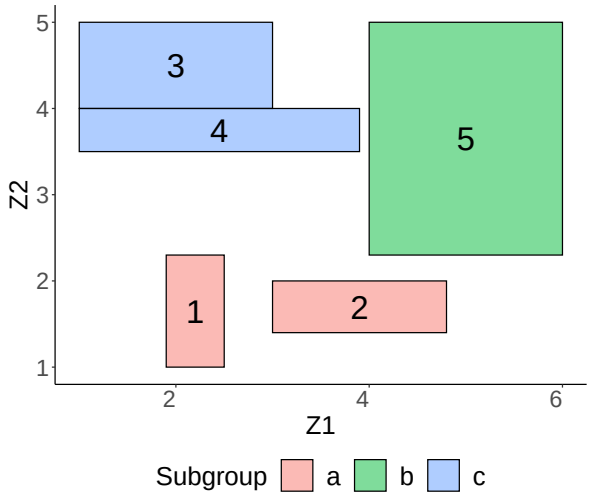


FIGURE 2.2: Hypothetical example of the shape of three $BM+$ groups denoted by a, b and c defined by two covariates Z_1 and Z_2. Three different shapes are illustrated which can be obtained by the considered methods. Subgroup a and c consisting of the union of the regions 1 and 2 or regions 3 and 4, respectively, can be obtained by all methods except ARDP and SIDES. Subgroup b consists of just one region and can be obtained by all methods. In practice the shape of subgroup b is preferable due to the lower number of thresholds which facilitates the assignment of subjects to the $BM+$ group.

Simulation study

The performance of the five methods in combination with the proposed subgroup criterion was evaluated by means of Monte Carlo simulations. In the following I

only present some selected results of the simulation study which, however, summarise the key findings. A detailed description of the chosen tuning parameters for the five methods, the model and parameter value assumptions used for generating the data, and other results can be found in Huber et al. [42].
The generated data sets consist of N subjects. For each subject a continuous outcome variable Y, a treatment indicator variable A and four covariates $Z_1, \ldots, Z_4$ are generated. The treatment indicator A is drawn from a binomial distribution $\mathcal{B}(1, 0.5)$ and the covariates $\mathbf{Z} = (Z_1, \ldots, Z_4)$ are drawn from $\mathcal{N}_4(0, \mathbf{I}_4)$ with $\mathbf{I}_4$ denoting the identity matrix. For generating the outcome $Y_i = \mu(A_i, \mathbf{Z}_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$ $(i = 1, \ldots, N)$ different mean functions $\mu(A, \mathbf{Z})$ are used. The parameter values used for the results are presented in Table 2.1 and are illustrated in Figure 2.3 and 2.4.

TABLE 2.1: Parameters in the simulation study for assessing the selection accuracy.

Parameter	Values
Model	M0: $\mu(A, \mathbf{Z}) = 0.2 \cdot A + \gamma \cdot I(Z_1 > 0)$ M1: $\mu(A, \mathbf{Z}) = 0.2 \cdot A + \gamma \cdot I(Z_1 > 0) + \beta_1 \cdot A \cdot I(Z_1 > 0)$
Sample size N	600, 1200, 2400
Main effect γ	$-0.2, 0, 0.2$
Interaction effect β_1	0.3 (small), 0.5 (medium), 1 (large)

For each parameter combination, 500 data sets were generated. For the subgroup criterion the threshold $\Delta_{\text{mintrt}} = 0.4$ was chosen. Based on this threshold the step function model M1 defines the $BM+$ subgroup $Z_1 > 0$ for all selected sizes of the interaction effect β_1. Therefore, the true treatment effect in the $BM+$ subgroup in settings with M1 is $\beta_1 + 0.2$. The true treatment effect in the complementary subgroup, the $BM-$ group defined by $Z \leq 0$, is 0.2. The null model, M0, does not include any treatment-by-biomarker interaction and is therefore used to evaluate the false discovery rate (FDR) of the methods in combination with the subgroup criterion. The FDR evaluates the relative frequency of identifying a $BM+$ subgroup although no subgroup is present in the data, i.e. the treatment effect is homogeneous across the entire population. It is considered that a $BM+$ subgroup does not include all or any patients. The FDR is sometimes also referred to as type I error rate, although strictly speaking no hypothesis is tested. Figure 2.3 shows that the erroneous identification of a target population, $BM+$, although the treatment effect is homogeneous across the population, occurs less frequently for IT and STIMA as compared to the other three methods. For MOB also smaller FDRs in comparison to SIDES and ARDP are observed. However, in presence of prognostic effects, i.e. $\gamma \neq 0$, the FDR of MOB is higher compared to settings without a prognostic effect as MOB evaluates instabilities in both the intercept and the treatment effect parameter of the underlying model (see Equation (2.1)). The FDR of SIDES decreases with increasing sample size and does not seem to be influenced by the presence of a prognostic effect. ARDP, however, does not include any FDR control and always results in the selection of a target population with the exception of two situations: (a) when the estimated treatment effect in the overall population is larger than the subgroup criterion threshold Δ_{mintrt} and (b) when none of the identified nested subgroups exceeds the threshold.

Figure 2.4 shows the relative frequency of not identifying a $BM+$ subgroup although a "true" target population, a $BM+$ group, is present. If the selected $BM+$

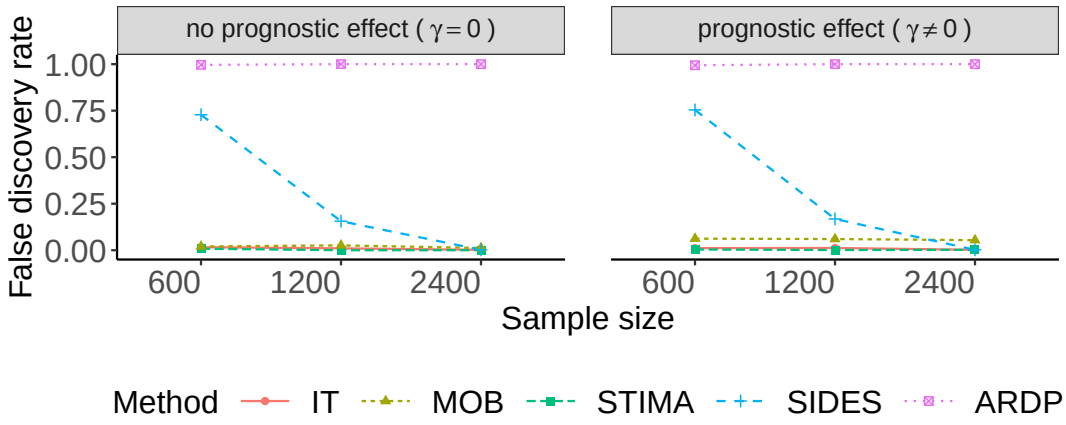


FIGURE 2.3: False discovery rate (FDR) for model M0 as mean function of the data generating model. The y-axes represent the FDR and the x-axes the considered sample sizes N. The left panel shows the FDR for setting without any prognostic effect, i.e. $\gamma = 0$ in M0 and the right panel shows settings with $\gamma = -0.2$ or $\gamma = 0.2$.

group includes all or any subjects, it is considered that no $BM+$ group was identified. This is also referred to as type II error corresponding to falsely retaining the null hypothesis of no subgroup being present. It is worth noting that here no hypothesis is tested, i.e. this definition of type II error has no connection to significance testing. The type II error rates illustrated in Figure 2.4 are based on model M1 without any prognostic effect, i.e. $\gamma = 0$. For smaller interaction effects all five methods have difficulties identifying a $BM+$ subgroup. For larger interaction effects STIMA and MOB show the lowest type II error rates. With increasing sample size the type II error rate of IT decreases. However, MOB outperforms the other methods clearly for medium-sized effects or smaller effects in combination with larger sample sizes. For ARDP, the smallest type II error rates are observable in settings with smaller interaction effects. However, this is due to the treatment effect in the overall population and the chosen value of the subgroup criterion Δ_{mintrt}. I further investigated how well the identified biomarker-positive $\widehat{BM+}$ subgroup and the identified biomarker-negative $\widehat{BM-}$ subgroup coincides with the true subgroups by means of additional performance criteria, e.g. the selection accuracy (P(subjects are correctly classified) with P denoting the probability) or the sensitivity (P(subjects are assigned to the $\widehat{BM+}$ group|subjects truly belong to the $BM+$ group)) and specificity (P(subjects are assigned to the $\widehat{BM-}$ group| subjects truly belong to the $BM-$ group)). For M1 as data generating model, MOB shows the best selection accuracy for even smaller treatment-by-covariate interactions and smaller sample sizes. However, the performance of MOB is influenced by the presence and direction of prognostic effects (results omitted).
Additional scenarios, not reported here, including a linear interaction trend, all splitting candidates being prognostic or a qualitative interaction were considered to evaluate their influence on the performance of the methods. The results of the additional scenarios and performance criteria can be found in Huber et al. [42]. For larger treatment-by-biomarker interaction effects and larger sample sizes, the approaches MOB, IT, and STIMA identify the $BM+$ subgroup similarly well with the exception of settings in which all potential splitting candidates $Z_1, \ldots, Z_5$ are prognostic. In these settings, STIMA performs best. For data with 600 subjects and smaller

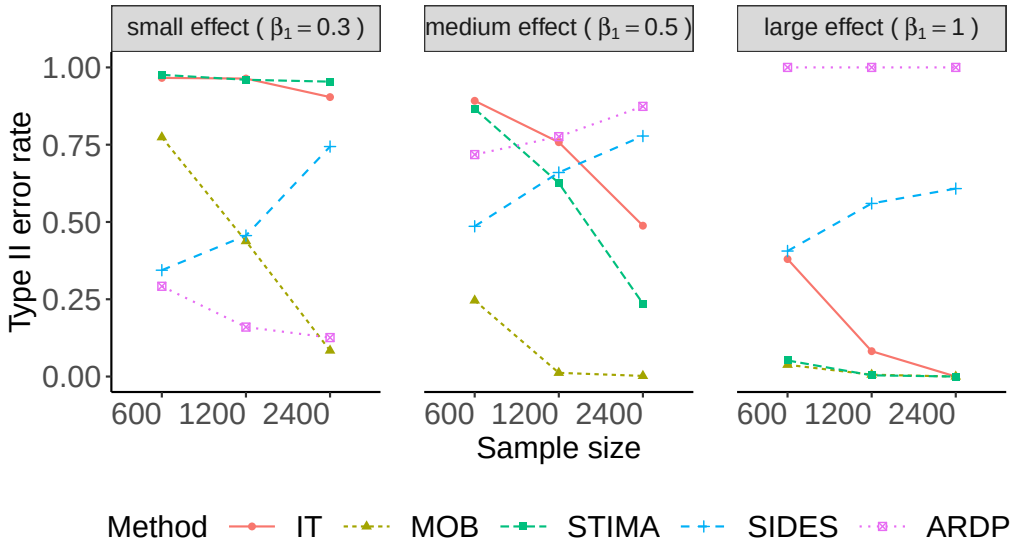


FIGURE 2.4: Type II error rate vs sample size for three interaction effect sizes β_1 in setting with the step function model M1 using $\gamma = 0$.

treatment-by-biomarker interactions, all methods have difficulties identifying a target subgroup. Overall, MOB seems the most promising method as it assigns the majority of patients correctly to $BM+$ and $BM-$ even in settings with fewer observations or smaller treatment-by-biomarker interactions.

Application example

For illustration purposes, I applied the five subgroup identification methods with the subgroup criterion to data from amyotrophic lateral sclerosis (ALS) patients, an orphan disease affecting the nervous system with a prevalence between 4.1 and 8.4 per 100 000 persons [60]. I used the PRO-ACT (Pooled Resource Open Access ALS Clinical trials) database aggregating data from 23 phase II/III trials due to the small sample sizes of single clinical studies on ALS [3]. The identified $\widehat{BM+}$ subgroup of MOB, IT, ARDP, and STIMA are similar. The results of MOB, IT, and ARDP only vary due to the cut-off value of the selected splitting variable. The $\widehat{BM+}$ subgroup identified by STIMA is more complex as it consists of the union of a subgroup similar to the one of MOB, IT, and ARDP and four smaller subgroups. The identified subgroup of MOB, IT and ARDP is defined by subjects with phosphorus values greater than 1.42 (1.36 for ARDP). Details on the application of the subgroup identification methods in combination with the proposed subgroup criterion to ALS data can be found in Section 4 of Huber et al. [42].

Summary

Summarizing, in my research on the comparison of subgroup identification methods, I presented five popular methods in a unified notation and proposed a criterion for selecting a target population based on a predefined minimum clinical benefit threshold. Using a simulation study I showed that MOB, STIMA and IT control the error of wrongly identifying a subgroup while the treatment effect is homogeneous across the entire population well. However, the type II error of mistakenly

retaining the null hypothesis of no subgroup being existent, is larger for all methods when the differential treatment effect or the sample size are smaller. Since real studies tend to have smaller sample sizes, it can be expected that the methods perform rather poorly in real-world scenarios. However, for the different scenarios considered, MOB showed the best performance.

2.3 Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning

In Subsection 1.2.3 I described the need for an approach combining commonly made assumptions in meta-analysis with a subgroup identification procedure. The GLMM-tree framework introduced by Fokkema et al. [32] uses the approach by Sela and Simonoff [78] for extending MOB, and therefore accounts for between-study heterogeneity. However, the authors only investigated a simpler model not accounting for heterogeneity in the treatment effect. Next, I introduce metaMOB which combines commonly made assumptions in meta-analysis with the GLMM-tree framework. The details of my research on metaMOB are published in Huber et al. [43].
For metaMOB data from $k = 1, \ldots, K$ randomized controlled trials investigating the same experimental treatment against the same control are assumed. The number of participants per trial is denoted by N_k. The outcome of participant i is denoted by y_i and a_i denotes the observed treatment group. It is assumed that each participant i is included in one trial only. The affiliation of each participant to one of the trials k is therefore already reflected in the index i and the index k is omitted to simplify the notation. Furthermore, the observed p baseline covariates for partitioning the data are denoted by $z_1, \ldots, z_p$.
The generalised linear models fitted by MOB to each resulting subgroup j are defined as

$$g(\mu_{ij}) = \gamma_j + \theta_j a_i \tag{M_0}$$

with μ_{ij} denoting the expected outcome of patient i in subgroup j and $g(\cdot)$ denotes a suitable link function. Fokkema et al. [32] considered heterogeneity in the baseline by using model (M_1):

$$g(\mu_{ij}) = \gamma_j + \theta_j a_i + b_{0k} \text{ with } b_{0k} \sim \mathcal{N}(0, \tau_0^2). \tag{M_1}$$

The between-study variance for the intercept is denoted by τ_0^2.
For subgroup identification via metaMOB different GLMMs are assumed and fitted to the identified subgroups. The differences in the models arise from the different assumptions regarding heterogeneity in the baseline. The GLMMs with a random baseline effect fitted in each of the subgroups j identified by metaMOB are defined in Equation (M_2):

$$\begin{aligned} &g(\mu_{ij}) = \gamma_j + b_{0k} + \theta_j a_i + b_{1k} a_i \\ &\text{with } b_{0k} \sim \mathcal{N}(0, \tau_0^2) \\ &\text{and } b_{1k} \sim \mathcal{N}(0, \tau_1^2). \end{aligned} \tag{M_2}$$

The random effects b_{0k} and b_{1k} are assumed to be normally distributed and, as commonly assumed, independent [48]. The between-study variance for the treatment is denoted by τ_1^2.

For metaMOB with stratified intercepts Equation (M_3) is fitted to each subgroup j:

$$g(\mu_{ij}) = \gamma_{jk} + \theta_j a_i + b_{1k} a_i \quad (M_3)$$
$$\text{with } b_{1k} \sim \mathcal{N}(0, \tau_1^2).$$

The baseline effect for trial k in subgroup j is denoted by γ_{jk} and is assumed to be fixed. Due to the subgroup and trial-specific intercept terms, the number of parameters to be estimated increases with the number of trials and identified subgroups. For both models (Equations (M_2) and (M_3)), the fixed effect part is assumed to be subgroup-specific, whereas the random effect part is assumed to be the same across the identified subgroups.
In the following, I refer to metaMOB with model (M_2) underlying the tree growing procedure as metaMOB-RI due to the **r**andom **i**ntercept. The method metaMOB with model (M_3), the **s**tratified **i**ntercept model, is referred to as metaMOB-SI. Model (M_1) does not consider between-study heterogeneity in the treatment effect and models between-study variation in the baseline by using a random intercept. Therefore, this method is referred to as MOB-RI.
The GLMMs (M_1) to (M_3) and (M_0) are special cases of the GLMM framework and can be represented by the model equation

$$g(\mu_{ij}) = \mathbf{x_i^T} \begin{pmatrix} \gamma_j \\ \boldsymbol{\theta_j} \end{pmatrix} + \mathbf{v}_i^T \boldsymbol{b}, \quad (2.5)$$

where $\mathbf{x_i}^T$ and $\mathbf{v}_i^T$ are the i-th row of the design matrix $\mathbf{X}$ for the fixed effects and $\mathbf{V}$ for the random effects, respectively. The vector $\mathbf{b}$ denotes the random effects, whereas the coefficient vector of the fixed effects is denoted by$(\gamma_j, \boldsymbol{\theta}_j)^T$. The vectors $\mathbf{b}$, $\mathbf{x}_i$ and $\mathbf{v}_i$ depend on the model ((M_0)-(M_3)) underlying the partitioning algorithm. The algorithm for partitioning the data based on the GLMMs (M_1)-(M_3) fits the chosen GLMM with the identified subgroups j as main effects to the data. The estimated random effects $\hat{b}$ are extracted from the estimated GLMM in order to use $\hat{b}$ as offset in the next step, i.e. the estimation of the tree based on the MOB algorithm [96]. The algorithm assumes either that the random effect part is known to estimate the fixed effect part via MOB or that the fixed effect part, the subgroup structure, is known to estimate the random effect part via fitting the GLMM. More details on the algorithm for MOB-RI, metaMOB-RI, and metaMOB-SI can be found in Huber et al. [43].
The methods MOB [77, 96] with (M_0), MOB-RI [32] with (M_1), metaMOB-RI with M_2 and metaMOB-SI with (M_3) as underlying model were compared in a simulation study considering different IPD settings.
The continuous response in the simulation study is generated by $y_i = f(z_i, a_i) + b_{0k} + b_{1k} \cdot a_i + \epsilon_i$, with $\epsilon_i \sim N(0, 5^2)$, $b_{0k} \sim N(0, \tau_0^2)$ and $b_{1k} \sim N(0, \tau_1^2)$. The generated data sets $\{(y_i, a_i, z_{i,1}, \ldots, z_{i,15}) : i = 1, \ldots, N\}$ consist of N subjects of equally sized trials ($k = 1, \ldots, K$). The treatment indicator A is drawn from a binomial distribution with a probability of 0.5. The 15 covariates Z_1,...,Z_{15} are drawn from a multivariate normal distribution with $\mu_{Z_1} = 10, \mu_{Z_2} = 30, \mu_{Z_4} = -40$ and $\mu_{Z_5} = 70$. The variance of all all covariates Z_p is set to $\sigma^2_{Z_p} = 100$ and the covariates are correlated with $\rho = 0.3$. The other means are drawn from a discrete uniform distribution on the interval $[-70, 70]$, following the data generating process described in Dusseldorp et al. [20] and Fokkema et al. [32].
The range of the parameter values considered for the simulation study are listed in Table 2.2. Additionally, the correlation between b_0 or b_1 and one of the potential splitting variables $\mathbf{Z}$ is varied: b_0 (or b_1) and all $\mathbf{Z} = (Z_1, \ldots, Z_p)$ uncorrelated, b_0

(or b_1) correlated with one of the splitting variables (correlation $\rho \approx 0.42$), b_0 (or b_1) correlated with one of the non-splitting variables (correlation $\rho \approx 0.42$). For each scenario 2000 data sets are generated.

TABLE 2.2: Scenarios considered in the simulation study.

Parameter	Values
Number of trials K	5, 10
Sample size N	200, 500, 1000
Heterogeneity in baseline τ_0	0, 5, 10
Heterogeneity in treatment τ_1	0, 2.5, 5, 10

When fitting GLMMs ((M_1),(M_2) or (M_3)) with the `lme4` R-package in the tree algorithm, convergence problems occur for both MOB-RI and metaMOB. The frequency of convergence problems using MOB-RI and metaMOB-SI was less than 0.8% across all the simulated data sets. The most convergence warnings, in around 1.7% of the simulated data sets, were obtained for metaMOB-RI based on model (M_2) which estimates the variance component of two random effects. In practice, the random effect structure is simplified when convergence problems are present, as the more complex random effect structure seems to drive convergence problems. Therefore, metaMOB-SI should be used instead of metaMOB-RI if convergence problems are encountered, as recommended by Kontopantelis [52] for IPD meta-analysis one-stage models. The calculation of further performance measures is based on cases without convergence warnings.

For assessing the FDR of the methods the fixed effect part in the data generating model is set to $f(\cdot) = 0$. In settings without correlation of the random effects b_0 and b_1 with one of the covariates, the false discovery rate is below 0.055 with a simulation error of approximately 0.49% for all the considered methods (not shown here). Figure 2.5 depicts the false discovery rate for settings with $K = 5$ and in which b_1 is correlated with one of the covariates $Z_1, \ldots, Z_p$. Without heterogeneity in both the baseline ($\tau_0 = 0$) and the treatment effect ($\tau_1 = 0$), all methods show similar FDRs across different sample sizes. With increasing heterogeneity in both baseline and treatment ((M_2) as the true underlying model), the FDR of both MOB and MOB-RI increases as both methods do not account for heterogeneity in the treatment effect. Since metaMOB-RI and metaMOB-SI account for these different types of heterogeneity, they are not strongly affected by their presence and do not show a worse FDR when heterogeneity is not present.

To assess the frequency of identifying the correct subgroups with differential treatment effects, data using the between-trial assumptions of (M_2) (not shown here) and (M_3) were generated. The outcome y_i according model (M_3) is generated by $y_i = f(z_i, a_i) + b_{1k} \cdot a_i + \epsilon_i$, with $\epsilon_i \sim N(0, 5^2)$, and $b_{1k} \sim N(0, \tau_1^2)$. For the fixed effect $f(\cdot)$ the tree structure presented in Figure 2.6 was used. The terminal nodes of the tree are the "true subgroups" which are defined by the covariates Z_1, Z_2, and Z_5. A tree was considered to be accurately estimated if it has the correct number of terminal nodes, all splitting variables are selected correctly and when the selected cut-off values for the split denoted by c are in the interval $c \pm 5$ with 5 corresponding to the population standard deviation. The greatest number of accurately estimated trees is observed for metaMOB-SI across all considered parameter variations, see Figure 2.7. However, the other three methods perform similarly well when there is no between trial heterogeneity in the treatment effect and when the random effects and the covariates $\mathbf{Z}$ are not correlated. As MOB and MOB-RI do not account for between-trial

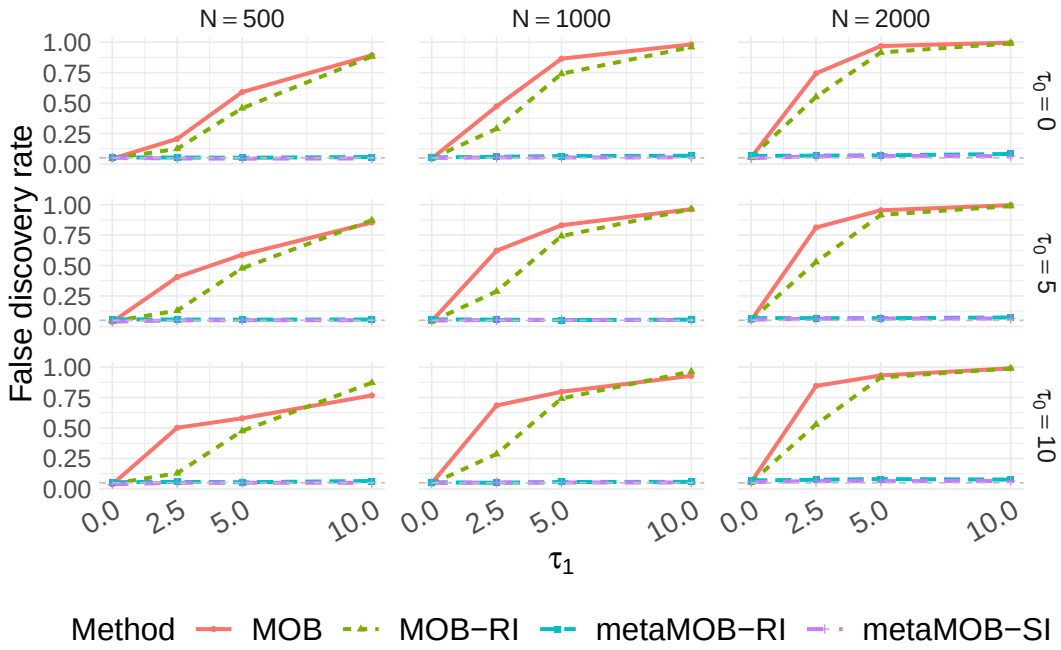


FIGURE 2.5: False discovery rate for settings with $K = 5$, $\text{cor}(b_1, \mathbf{Z}) \neq 0$ and $\text{cor}(b_0, \mathbf{Z}) = 0$. The dotted line at value 0.05 indicates the prespecified level of significance used as stopping criteria in the algorithm.

heterogeneity in the treatment effect, fewer trees are accurately estimated with increasing variance of the random treatment effect τ_1^2 compared to metaMOB-SI and metaMOB-RI if random treatment effects b_1 and a splitting variable Z_{split} (either Z_1, Z_2 or Z_5) are correlated. The assumption for the heterogeneity in the baseline of metaMOB-RI is not flexible enough for settings with trial and subgroup-specific intercepts ($\tau_\gamma \neq 0$). Based on the underlying model of metaMOB-RI ((M_2)), heterogeneity in the baseline effects is assumed to be constant across studies, which is not the case for the data-generating model of the second row of Figure 2.7. Therefore, its tree accuracy decreases with larger variations of the trial and subgroup-specific intercepts τ_γ.

To assess the performance of the methods regarding treatment effect estimation in the estimated subgroups, the correlation between the estimated and true treatment effect was calculated. Throughout the considered settings with (M_2) and (M_3) as data-generating models, the more complex models metaMOB-RI and metaMOB-SI showed the highest correlation of true and estimated treatment effect in the estimated subgroups (not shown here, see Figure 6 in Huber et al. [43]).

Summarizing, I proposed the methods metaMOB-SI and metaMOB-RI, which fall into the broad class of GLMM-trees [32] for subgroup identification in IPD meta-analysis. I showed that accounting for heterogeneity in settings in which no between-trial heterogeneity is present or when between-trial heterogeneity is independent of potential splitting candidates is less relevant due to the similar performance of the considered methods. However, as the composition of the trial populations in terms of patient characteristics contributes to the heterogeneity between trials, it is reasonable to assume that the heterogeneity between trials is linked to one or more patient-level covariates. By utilizing Monte Carlo simulations, I demonstrated that the misspecification of the between-trial heterogeneity structure in settings in which

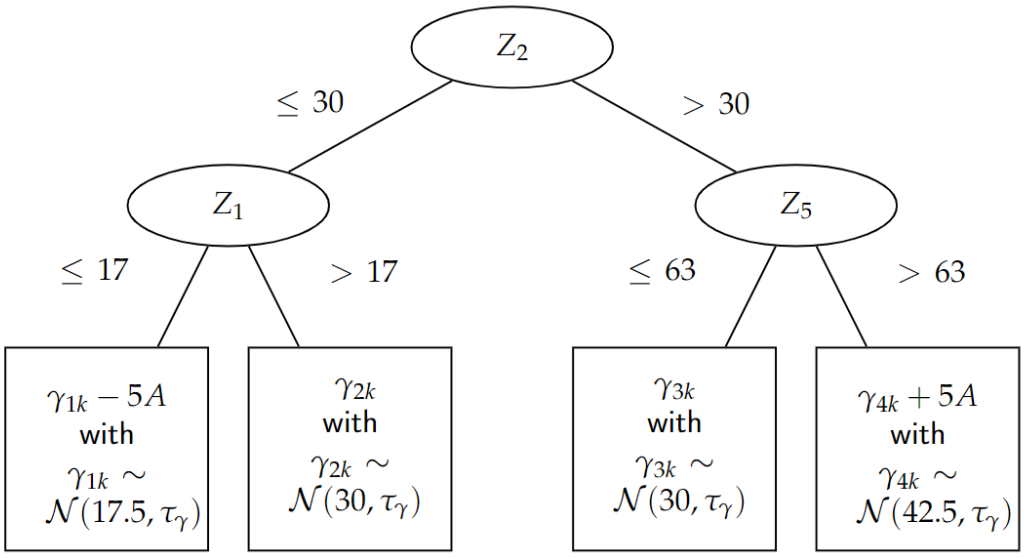


FIGURE 2.6: Fixed effect structure $f(\cdot)$ for the data generating model. Z_1,Z_2 and Z_5 denote the covariates defining the four subgroups. The true fixed intercepts are denoted by γ which are drawn from a normal distribution.

covariates and heterogeneity are linked, adversely affects the FDR, the accuracy of the estimated tree, and the estimated treatment effect for the identified subgroups. I concluded that metaMOB-SI is the preferred option for subgroup identification in IPD meta-analysis as it showed the best performance across the considered settings. Furthermore, the underlying model accounts for between-trial heterogeneity in the treatment effect and models baseline heterogeneity with fixed effects making it the more flexible approach than metaMOB-RI, which imposes a constraint on the baseline effects. However, the number of parameters that have to be estimated for the underlying models of metaMOB-SI increases with the number of trials and identified subgroups which might lead to inconsistent estimators as described by the Neyman–Scott problem [65].

2.4 Model-based recursive partitioning for discrete event times

For analysing data with a time-to-event outcome measured on a discrete time scale appropriate statistical methods are needed as introduced in Section 1.2.4. In Huber et al. [45] I proposed a method for identifying subgroups defined by prognostic and/or predictive biomarkers tailored to a discrete time-to-event outcome based on the MOB algorithm.

Parametric regression models for discrete event times are usually based on the *discrete hazard function*, which has the form $\lambda(t|\mathbf{X}) = P(T = t|T \geq t, X = \mathbf{x})$, describing the conditional probability of the event at time point t given survival until t [89]. The discrete event time T can take values in $\{1, \ldots, L\}$, which might result from L underlying intervals $[a_0, a_1), [a_1, a_2), \ldots, [a_{L-2}, a_{L-1}), [a_{L-1}, \infty)$. The parametric discrete hazard model is defined by

$$g(\lambda(t|\mathbf{x})) = \gamma_{0t} + \mathbf{x}^T\boldsymbol{\beta}, \tag{2.6}$$

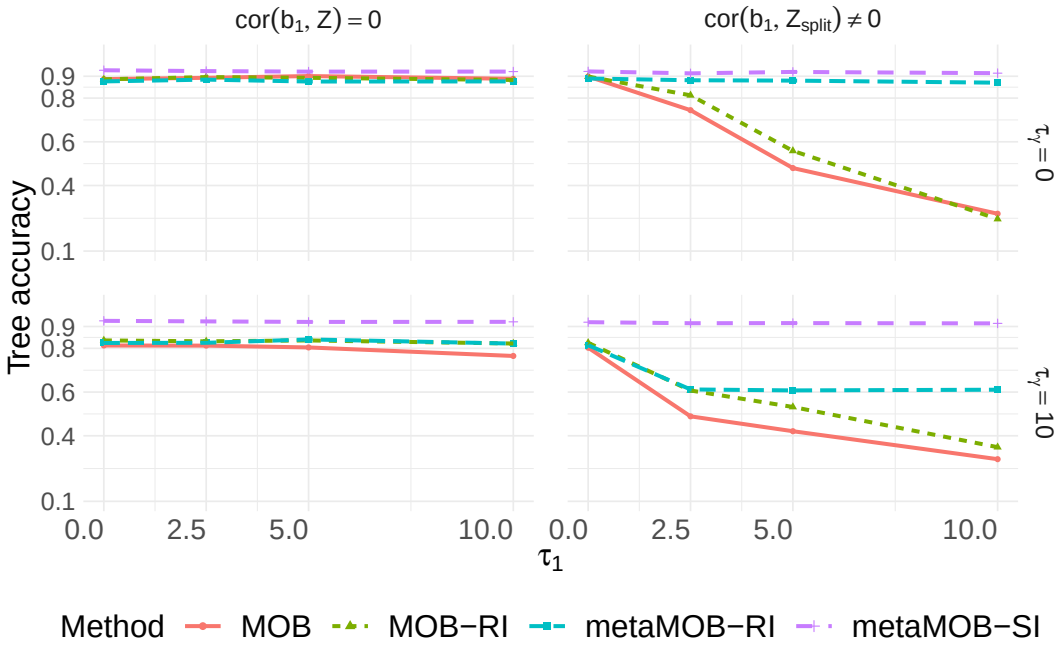


FIGURE 2.7: Tree accuracy for model (M_3) as data-generating mechanism. The correlation of b_1 and the covariates $\mathbf{Z}$ are varied (columns). Different variances for the subgroup and trial specific intercepts are presented in the three rows.

where $g(\cdot)$ is a monotonic link function relating conditional survival probabilities to the covariates $\mathbf{x}$ by a vector of regression coefficients $\boldsymbol{\beta}$. The set $\boldsymbol{\gamma_0} = \{\gamma_{01}, \ldots, \gamma_{0,L-1}\}$ defines a covariate-free "baseline" trend (for $\mathbf{x} = 0$). Common discrete time-to-event models are the proportional continuation ratio models using the logistic function for $h(\eta) = g^{-1}(\eta) = \exp(\eta)/(1 + \exp(\eta))$ or the grouped proportional odds model using the Gompertz distribution $h(\eta) = 1 - \exp(-\exp(\eta))$.
As the likelihood of discrete time-to-event models (Equation (2.6)) is equivalent to the likelihood of a binomial model that distinguishes whether the event occurred at time t or not (given $T \geq t$), the generalized linear modelling (GLM) framework can be used for modelling discrete hazards $\lambda(t|\mathbf{x})$ [76, 89]. For fitting discrete hazard models, the binary response y_{it} of subject i at time $t = 1, \ldots, \tilde{T}_i$, is defined as

$$y_{it} = \begin{cases} 1 \text{ if } t = \tilde{T}_i \text{ and } \delta_i = 1, \\ 0 \text{ else,} \end{cases} \tag{2.7}$$

with $\delta := I(T \leq C)$ denoting the status indicator indicating whether the observed survival time $\tilde{T}$ is right-censored ($\delta = 0$) or not ($\delta = 1$). From the data matrix $\mathcal{D} = \{(\tilde{t}_i, \delta_i, \mathbf{x}_i^T, \mathbf{z}_i^T) : i = 1, \ldots, N\}$, the augmented data matrix with the binary response for fitting the discrete hazard models is constructed: $\mathcal{D}_A = \{(y_m, t_m^*, \mathbf{x}_m^{*T}, \mathbf{z}_m^{*T}) : m = 1, \ldots, n\}$. Alongside the binary response y, the augmented data matrix consists of a time index for each subject i given by $(t_1^*, \ldots, t_N^*)^\top$ with $t_i^* = (1, 2, \ldots, \tilde{T}_i)$, $i = 1, \ldots, N$ allowing to fit time-dependent intercepts. Additionally, two disjoint sets of covariates $\mathbf{x}$ and $\mathbf{z}$ are considered. The covariates $\mathbf{x}$ are used to fit the regression model in the nodes of MOB, whereas the covariates $\mathbf{z}$ are potential splitting candidates. The covariate vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ are duplicated subject-wise for the augmented matrix. The subject-wise duplicated vectors are denoted by $\mathbf{x}_i^*$ and $\mathbf{z}_i^*$, respectively. The number of rows in the augmented data matrix for subject i is $\tilde{T}_i$. Therefore, the

number of rows in the augmented data matrix is $n = \sum_i \tilde{T}_i$. Although the data sets $\mathcal{D}$ and $\mathcal{D}_A$ differ in the number of rows, the number of columns is the same despite different contents.
Although fitting discrete time-to-event models is based on the GLM framework which is used in MOB as well, MOB cannot be directly applied to discrete time-to-event data. Using the augmented data matrix to fit the discrete hazard model, which is assumed to be the underlying model of MOB for time-to-event data, results in an increased type I error rate of the M-fluctuation test as I demonstrated in Huber et al. [45]. By means of a Monte Carlo simulation study, I demonstrated that ignoring dependencies in the augmented data matrix leads to a systematic inflation of the type I error rate in case the standard MOB approach with binary outcome is applied to a set of discrete time-to-event data. Furthermore, I proposed MOB for discrete survival outcomes (MOB-dS) an adjusted MOB algorithm tailored to model discrete time-to-event data. The adjusted algorithm is based on a permutation approach for obtaining the distribution under the null hypothesis of the M-fluctuation test used for partitioning the data using MOB-dS.
For partitioning the data the M-fluctuation test [96, 95] is applied to each resulting node j. Without restriction of generality, the root node is considered in the following and therefore the index j is omitted to enhance readability. The M-fluctuation test investigates instabilities in the model parameters γ and $\boldsymbol{\beta}$ which are estimated by maximizing the log-likelihood $\Psi(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$. The null hypothesis for variable Z_r ($r = 1 \ldots, p$) in the M-fluctuation test is

$$H_0^{\theta,r} : \boldsymbol{\theta}_k = \boldsymbol{\theta}_0, \quad k = 1, \ldots, n \tag{2.8}$$

versus the alternative that (at least one component of) $\boldsymbol{\theta}_k$ varies across Z_r . The parameter vector $\boldsymbol{\theta}_k$ ($k = 1, \ldots, n$) is a row-specific vector of regression coefficients based on $\mathcal{D}_A$. The test statistic of the M-fluctuation test is based on the empirical fluctuation process $W_r(o)$ (see Equation (2.9)) and a scalar function $\zeta(\cdot)$, e.g. the supLM statistics [2], which is applied to the empirical fluctuation process. This results in $\zeta(W_r(\cdot))$ and the corresponding limiting distribution being $\zeta(W^0(\cdot))$ with the empirical fluctuation process converging to the Brownian bridge denoted by W^0. The empirical fluctuation test is defined by

$$W_r(o) = \hat{J}^{-1/2} n^{-1/2} \sum_{u=1}^{\lfloor no \rfloor} h'(\mathbf{x}_u^* \hat{\theta}) \frac{y_u - \hat{\mu}_u}{\hat{\mu}_u(1 - \hat{\mu}_u)} \mathbf{x}_u^*, \quad 0 \leq o \leq 1 \tag{2.9}$$

with $\mathbf{x}_u^*$ denoting row u of the augmented data matrix $\mathcal{D}_A$ ordered by the values of Z_r. The floor function is denoted by $\lfloor \cdot \rfloor$. The empirical fluctuation process is the partial sum process of the scores ordered by the variable Z_r, scaled by the number of the rows n and $\hat{J}$, an estimate of the covariance matrix $\text{Cov}(\psi(Y, X, \hat{\theta}))$. The score function corresponding to the log-likelihood is denoted by $\psi(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \frac{\partial \Psi(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\partial \theta}$. Depending on the scale of Z_r a corresponding scalar function for defining the test statistic is chosen, i.e. supLM statistics [2] for numerical Z_r or the weighted sum of the squared L_2 norm of the increments of the empirical fluctuation process over the observations in one of the categories of Z_r for categorical Z_r. Zeileis and Hornik [95] provide a detailed description of the M-fluctuation test.
In contrast to MOB which calculates the p-values using [38] or [40] depending on the scale of covariates $\mathbf{Z_1}, \ldots, \mathbf{Z_p}$, I propose to calculate the p-values based on the empirical distribution of the test statistic obtained by the following permutation procedure:

1 Randomly permute the row vectors $(\tilde{t}_i, \delta_i, \mathbf{x}_i^T)$ of the un-augmented data $\mathcal{D}$ against the covariate vector $\mathbf{z}_i^T$.

2 Form the augmented data of the permuted data.

3 Calculate the M-fluctuation test statistic based on the augmented data of the permuted data.

Step 1 of the permutation approach retains the overall effects of $\mathbf{X}$ and the correlations among $\mathbf{Z}$ while removing any marginal effects of the covariates $\mathbf{Z}$ on the outcome. The algorithm of MOB-dS differs only from MOB's algorithm for binomial data in the calculation step of the p-value: MOB-dS uses a permutation approach which accounts for the dependencies in the augmented data matrix. The permutation approach provided in the implementation of MOB in the R-package `partykit` is not appropriate for discrete time-to-event data. This approach permutes the rows of the data matrix used for fitting the underlying model, i.e. the augmented data matrix $\mathcal{D}_A$ for discrete time-to-event data. Since the common assumption regarding independent observations is violated [95, 96] for discrete time-to-event data due to the use of the augmented data matrix, the asymptotic theory of the M-fluctuation test used in MOB is not valid.
The type I error rate of the M-fluctuation test with the newly proposed permutation strategy for time-to-event data used in MOB-dS was determined by means of a Monte Carlo simulations and compared to the type I error rate of the approach by Zeileis and Hornik [95] used in MOB. The setup and results of the Monte Carlo simulation studies are summarized next.
The time-to-event T was generated using a proportional continuation ratio model (PCRM) with time-dependent baseline coefficients (Equation (2.10)) which were constant across the covariate values of $\mathbf{Z}$:

$$\lambda(t) = \frac{\exp(\gamma_{0t})}{1 + \exp(\gamma_{0t})}, \quad t = 1, \ldots, L-1. \tag{2.10}$$

The values of $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{L-1})$ were chosen in order to result in different shapes of survival functions, e.g. events occurring in the middle of the observational period. Furthermore, different event rates (20%, 40%, and 60%) were considered. The shape of the survival functions with the three different event rates used for generating the data is illustrated exemplarily for $L = 7$ in Figure 2.8. The true censoring times C were drawn from a continuous exponential distribution and were independent of the event time T. The observed true survival time $\tilde{T} = \min(T, C)$ and the status indicator $\delta := I(T \leq C)$ are obtained based on the event and censoring time. The number of discrete time points was varied $L \in \{4, 5, \ldots, 11\}$. The rates of the exponential distribution were chosen to achieve approximately 0%, 20%, and 50% censoring. Thirteen covariates Z_1,...,Z_{13} were drawn from a standard normal distribution with $\rho = 0.1$ being the correlation of the covariates. No additional variables $\mathbf{X}$ were considered in the data-generating model. The number of subjects was set to $N = 100 \cdot (L-1)$. For each scenario, 2000 Monte Carlo replications were generated. For both MOB and MOB-dS the PCRM defined in Equation (2.10) was considered as the underlying model. The sandwich estimator was used as the covariance matrix estimator in the parameter instability tests (see Equation (2.9)). For all other arguments of MOB and also for MOB-dS the default values of the R-package `partykit` version 1.2-15 were used. For MOB-dS 1000 repetitions of the permutation approach were conducted.
Figure 2.9 shows the inflated type I error rate of MOB across all the scenarios for

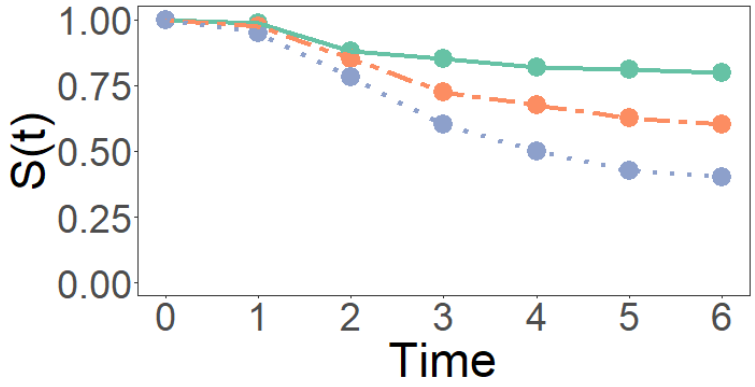


FIGURE 2.8: Illustration of the true survival functions for settings with $L = 7$. The different colours and line types correspond to settings with 80% (solid), 60% (dashed) and 40% (dotted) events.

which MOB-dS (with 1000 samples) controlled the type I error rate well. Fewer discrete time points, e.g. $L = 4$, and higher event rates, e.g. 60%, lead to fewer dependencies in the augmented data matrix. This might contribute to a type I error rate of MOB close to 5% in this setting. This is not only observed in settings with events occurring in the middle of the observational period as shown in Figure 2.9, but also in settings with linear survival functions and with events occurring mainly at the beginning of the observational period (results not reported here). Higher type I error rates for MOB are observable in settings with a larger number of discrete time points and fewer events.

To illustrate the difference of the approximate asymptotic and proposed sampling distribution of the M-fluctution test for discrete-time-to-event data and binary data, the exceedence probabilities are shown in Figure 2.10. The exceedence probabilities for a range of values of the test statistics used in MOB (black) and MOB-dS (red) likely to be of interest when testing are included in Figure 2.10. The distribution of MOB is based on the Hansen approximation [38] and the distribution obtained by MOB-dS is based on 1000 permutation samples. The right-hand tail probabilities for the test statistic of two covariates are shown for simulated data with $L = 8$ and 20% events, which resulted in higher type I error rates. The first row shows the obtained tail probabilities of the test statistic based on the proportional continuation ratio model (Equation (2.10)). The second row corresponds to the tail probabilities for test statistics based on logistic regression models ignoring the time-to-event information by using the status indicator δ as outcome. For binary outcomes both approaches lead to tail probabilities of the test statistic that are very close (see the second row of Figure 2.10). This is consistent with the type I error rate of MOB not being inflated for binary outcomes with independent observations. For discrete hazards models based on the augmented data matrix, however, the distribution obtained by MOB seems to be shifted compared to the one obtained by MOB-dS (see the first row of Figure 2.10), leading to the test used in MOB becoming anti-conservative. This confirms the results of the inflated type I error rate of the previous simulations. Thus, for discrete hazard models that use an augmented data matrix to fit a logistic model, the Hansen approximation does not seem to be suitable for approximating the asymptotic distribution of the instability test statistic used as splitting criterion in MOB.

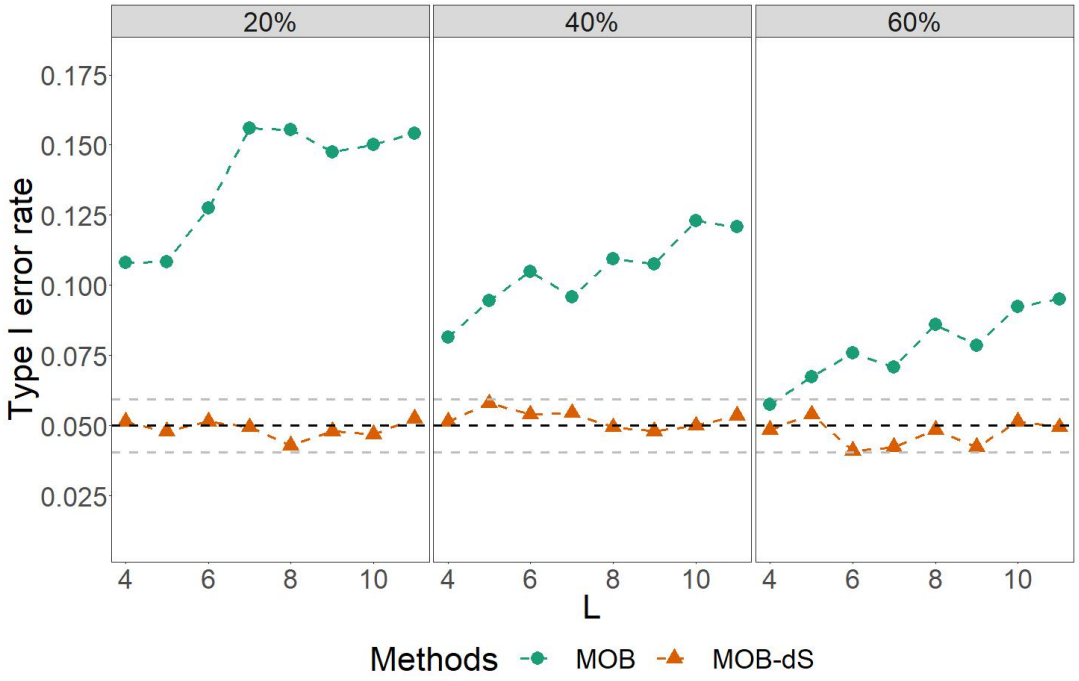


FIGURE 2.9: Type I error rate for MOB and MOB-dS based on 2000 simulated data sets per setting with 20% censoring and 1000 permutation samples for MOB-ds. The columns represent settings with the three different expected event rates. The horizontal grey lines mark $\alpha \pm 2SE$ with $SE = \sqrt{\frac{0.05 \cdot 0.95}{2000}} \approx 0.49\%$ being the Monte Carlo simulation error at a simulated type I error of 0.05.

Summarizing, I showed that although discrete time-to-event data can be modelled using the GLM framework, which is integrated into MOB, MOB cannot be directly applied to discrete time-to-event data. Applying MOB to discrete time-to-event data results in inflated type I error rates of the M-fluctuation test used as MOB's splitting criterion. Therefore, I proposed MOB-dS, a permutation approach, tailored to discrete time-to-event data. Furthermore, I showed via simulations that the type I error rate is controlled better by MOB-dS compared to MOB applied to discrete time-to-event data. However, the permutation approach used in MOB-dS is computationally expensive, because a sufficiently large number of permutations is required to approximate the distribution of the test statistic.

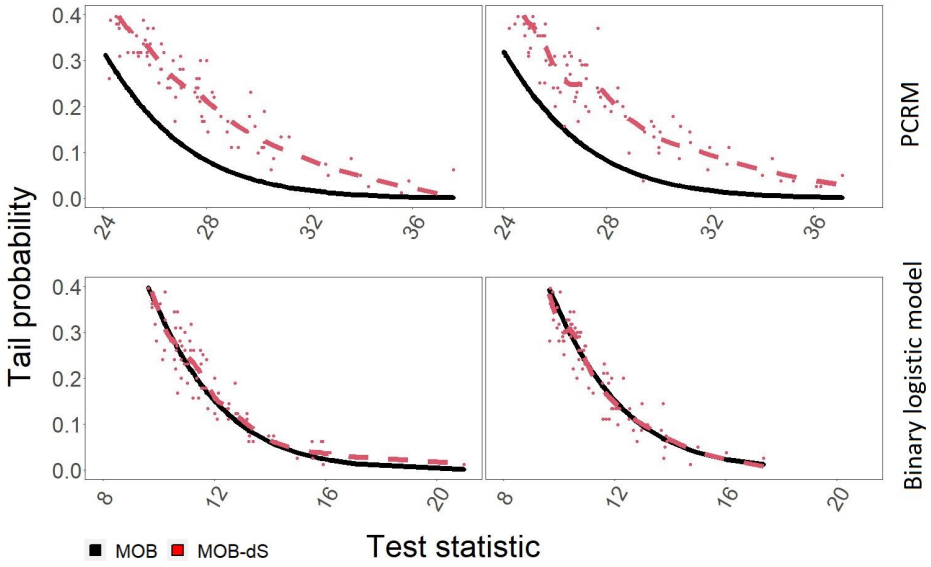


FIGURE 2.10: Right-hand tail probabilities from approximate asymptotic and sampling distributions of the instability test in MOB (solid black) and MOB-dS (dashed red). The dashed red lines correspond to the smoothed sampling distributions obtained by the LOESS estimator. Test statistics for two out of thirteen covariates are shown only. Distributions are illustrated for data generated with $L = 8$ and 20% events. The sampling distribution used in MOB-dS is based on 1000 permutation samples.

Chapter 3

Discussion

My research focused on subgroup identification in the context of randomized controlled trials in precision medicine. Biomarker-defined subgroups in drug development may be selected for different reasons and based on different sources of evidence. For assessing whether the presented evidence is acceptable for a biomarker-defined subgroup related to a specific drug, I proposed a classification scheme distinguishing between biological and data-driven evidence related to subgrouping [44]. In my dissertation, I proposed different approaches for the data-driven identification of subgroups with differential treatment effects based on data from multiple clinical trials or data with a discrete time-to-event outcome. The methods I proposed are based on model-based recursive partitioning, a subgroup identification method, that performed best in the neutral comparison study, investigating the performance of five methods for identifying biomarker-defined subgroups with differential treatment effect [42]. The advantages of the metaMOB methods are that the heterogeneity in baseline and treatment effect can be modelled appropriately which reduces false findings, i.e. identification of subgroups although the treatment effect is homogeneous across the population and subgroups defined by spurious biomarkers and cut-off values [43]. The benefit of the proposed change in MOB tailored to discrete time-to-event data is that it ensures controlling the type I error rate of the M-fluctuation test used as splitting criterion [45].
My research in Huber et al. [42] focused on comparing methods that identify a biomarker-positive subgroup. The biomarker-positive subgroup is defined by a better treatment effect than its complementary counterpart, namely the biomarker-negative subgroup. To obtain these biomarker-positive and biomarker-negative subgroups based on the results of different subgroup identification methods, I proposed to amalgamate subgroups identified by the considered methods whose estimated treatment effect exceeds the proposed subgroup criterion. However, the subgroup criterion could have been chosen differently, e.g. selecting the subgroup with the largest test statistic for the treatment effect as in Johnston et al. [50]. Alternatively, Ballarini et al. [5] proposed to use the lower or upper bound of confidence intervals for the treatment effect to identify a subgroup with a larger treatment benefit. However, inference after model selection remains a challenging task as the additional uncertainty from model selection has to be integrated into the inferential process. Post-selection inference has primarily been applied to regression models using the least absolute shrinkage and selection operator (lasso) [55, 88]. For other subgroup identification methods, e.g. the methods investigated in my comparison study [42], the application of post-selection inference [4, 8] remains a topic of future research.
The identification of subgroups and neutral comparisons of the different proposed methods is still a relevant topic as publications in the past few years [1, 19, 42, 59, 79] and the recent publication by Sun et al. [84] demonstrate. Although all studies focus on the comparison of subgroup identification methods, the results differ because of

the study-specific properties of the investigated data, the included methods, and the underlying scientific questions of interest. Strobl and Leisch [82] advocate a more differentiated view of the research question “Which is the best method in general", which they consider ill-posed. The performance on a single task depends on both the method used and the dataset properties. For instance, Loh et al. [59] investigate subgroup identification for binary endpoints, whereas Alemayehu et al. [1] and Huber et al. [42] consider continuous endpoints. Additionally, the focus of research questions differs, e.g. Sun et al. [84] focus on a more general assessment of the treatment heterogeneity, whereas in my research I focused on the identification of a biomarker-positive and biomarker-negative subgroup with differential treatment effects.
My research on subgroup identification based on IPD meta-analysis considers data from multiple trials [43]. My work showed that accounting for between-trial heterogeneity is essential for identifying accurate subgroups. Additionally, I recommend modelling between-trial heterogeneity in the treatment effect using random effects, and modelling the between-trial heterogeneity in the baseline using trial-specific fixed effects, as this approach has been shown to be robust in my simulations. A common assumption in meta-analysis is the normal distribution of random effects. This distribution assumption is also made almost automatically in linear mixed models, which are also used for IPD meta-analyses. A normal distribution was also assumed for the random effects in Huber et al. [43]. Misspecification of the shape of the random effects distribution in GLMMs has been shown to affect the random effects predictions, but has a smaller effect on the fixed effects estimates [17, 64]. Since the data partitioning of metaMOB is mainly based on the estimation of fixed effects, I assume that a misspecification of the random effects distribution should not have a larger impact on the tree structure, i.e. the selected splitting variables. However, this needs to be investigated further. Heterogeneity in treatment effects and baseline may also occur with other cluster structures in the data, e.g. data from different centres. The ICH E9 guideline [47] also states that the heterogeneity of the treatment effect between centres should be explored if positive treatment effects were found in the study. The proposed metaMOB approach can also be applied to data with heterogeneity that does not result from studies, but rather from other clustering structures. One assumption in metaMOB is that the subgroup structure is the same in the different clusters. However, it is plausible that the true subgroup structure varies slightly from study to study, e.g. the cut-off values of the biomarkers. In more extreme cases, the biomarkers that define the subgroups could also differ. Pooling data with different subgroup definitions will complicate identification, as differential treatment effects between subgroups may be obscured by the different subgroup definitions. The extent to which heterogeneity in subgroup definitions between clusters affects the performance of metaMOB, whether and how this heterogeneity is dealt with in the course of subgroup identification, is a topic for future research.
A more accurate subgroup identification may improve future trials in terms of the trial design, subgroup analysis plan, and even the decision on the need for further trials. More accurate subgroup findings in terms of subgroup definitions and treatment effect estimates are more likely to be confirmed in subsequent trials. For instance, sample size calculations and decisions on conducting further trials depend amongst others on the treatment effect estimates in the identified subgroups. Hence, controlling for overoptimistic treatment effect estimates improves the results of subgroup identification methods. A flexible and effective way to incorporate results of subgroup identification methods into the design and analysis plan are adaptive designs. Research on adaptive designs offering the possibility to select promising

subgroups and reallocate sample size was published recently [37, 71]. However, these approaches need prespecified subgroups, which might be based on the knowledge of the drug's mechanism of action or analyses of previous studies. Recently a proposal on a trial design incorporating subgroup identification and confirmation together in a framework was made. Johnston et al. [50] proposed a two-stage adaptive clinical trial design with data-driven subgroup identification at interim analysis. Four different data-driven subgroup identification methods were considered for the interim analysis, a brute force method, SIDES [58] and SIDEScreen adaptive [56] and lasso [88]. At the interim analysis, conditional powers based on the subgroups and the overall populations are used to decide whether to continue the study as planned, to increase the sample size in the overall population or the subgroup, to select the overall population, the subgroup or both for the final analysis, or to stop the study because of futility. The final analysis involves combination tests together with closed testing procedures. Due to the lack of comparison studies for subgroup identification methods for time-to-event data, Johnston et al. [50] selected the methods based on their popularity and the ease of their implementation. Simulation studies comparing the performance of different methods for selecting a target population as in my work [42] would be useful in the context of adaptive design concepts since more information on the method performance can be used to determine the subgroup identification method that should be incorporated in the interim analysis.
Precision medicine encompasses both differential treatment benefit and safety. Identifying patients at increased risk of adverse events is more challenging, as adverse events are sometimes rare. Furthermore, data on adverse events are usually accompanied by additional information, e.g. the number of occurrences, severity, timing and their duration [70]. This information should be taken into account in the analyses. Therefore, more complex methods are needed to analyse safety data. For instance, multi-state models may be required, because the occurrence of a particular adverse event may be precluded by the occurrence of another adverse event or by the occurrence of the primary time-to-event endpoint. Although I focused on treatment efficacy endpoints, the proposed methods apply in principle to safety outcomes as well.

Bibliography

[1] Alemayehu, D., Chen, Y., and Markatou, M. "A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations". In: *Statistical Methods in Medical Research* 27.12 (2018), pp. 3658–3678.

[2] Andrews, D. W. K. "Tests for Parameter Instability and Structural Change With Unknown Change Point". In: *Econometrica* 61.4 (1993), pp. 821–856.

[3] Atassi, N., Berry, J., Shui, A., Zach, N., Sherman, A., Sinani, E., Walker, J., Katsovskiy, I., Schoenfeld, D., Cudkowicz, M., and Leitner, M. "The PRO-ACT database Design, initial analyses, and predictive features". In: *Neurology* 83.19 (2014), pp. 1719–1725.

[4] Bachoc, F., Leeb, H., and Pötscher, B. M. "Valid confidence intervals for post-model-selection predictors". In: *The Annals of Statistics* 47.3 (2019).

[5] Ballarini, N. M., Rosenkranz, G. K., Jaki, T., König, F., and Posch, M. "Subgroup identification in clinical trials via the predicted individual treatment effect". In: *PloS One* 13.10 (2018), e0205971.

[6] Barnett, A. G., Batra, R., Graves, N., Edgeworth, J., Robotham, J., and Cooper, B. "Using a longitudinal model to estimate the effect of methicillin-resistant Staphylococcus aureus infection on length of stay in an intensive care unit". In: *American Journal of Epidemiology* 170.9 (2009), pp. 1186–1194.

[7] Bauer, P., Bretz, F., Dragalin, V., König, F., and Wassmer, G. "Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls". In: *Statistics in Medicine* 35.3 (2016), pp. 325–347.

[8] Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. "Valid post-selection inference". In: *The Annals of Statistics* 41.2 (2013).

[9] "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework". In: *Clinical Pharmacology and Therapeutics* 69.3 (2001), pp. 89–95.

[10] Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. "Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data". In: *Statistical Modelling* 11.5 (2011), pp. 429–446.

[11] Boulesteix, A.-L., Binder, H., Abrahamowicz, M., and Sauerbrei, W. "On the necessity and design of studies comparing statistical methods". In: *Biometrical Journal* 60.1 (2018), pp. 216–218.

[12] Boulesteix, A.-L., Lauer, S., and Eugster, M. J. A. "A plea for neutral comparison studies in computational sciences". In: *PloS one* 8.4 (2013), e61562.

[13] Boulesteix, A.-L., Wilson, R., and Hapfelmeier, A. "Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies". In: *BMC medical research methodology* 17.1 (2017), p. 138.

[14] Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. "Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology". In: *Statistics in Medicine* 28.10 (2009), pp. 1445–1463.

[15] Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and Regression Trees*. Monterey,CA: Wadsworth and Brooks, 1984.

[16] Brookes, S. T., Whitely, E., Egger, M., Smith, G. D., Mulheran, P. A., and Peters, T. J. "Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test". In: *Journal of Clinical Epidemiology* 57.3 (2004), pp. 229–236.

[17] Butler, S. M. and Louis, T. A. "Random effects models with non-parametric priors". In: *Statistics in medicine* 11.14-15 (1992), pp. 1981–2000.

[18] Cuijpers, P., van Straten, A., and Warmerdam, L. "Behavioral activation treatments of depression: a meta-analysis". In: *Clinical Psychology Review* 27.3 (2007), pp. 318–326.

[19] Doove, L., Dusseldorp, E., Van Deun, K., and Van Mechelen, I. "A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions". In: *Advances in Data Analysis and Classification* 8.4 (2014), pp. 403–425.

[20] Dusseldorp, E., Conversano, C., and Van Os, B. J. "Combining an Additive and Tree-Based Regression Model Simultaneously: STIMA". In: *Journal of Computational and Graphical Statistics* 19.3 (2010), pp. 514–530.

[21] European Medicines Agency. "CHMP/EWP/2459/02 - Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design". In: (2007).

[22] European Medicines Agency. "CHMP/EWP/908/99 - Points to consider on multiplicity issues in clinical trials". In: (2002).

[23] European Medicines Agency. "EMA/446337/2011 - Reflection paper on methodological issues associated with pharmacogenomic biomarkers in relation to clinical development and patient selection". In: (2011).

[24] European Medicines Agency. "EMA/CHMP/539146/2013 - Investigation of subgroups in confirmatory clinical trials - Scientific guideline". In: (2019).

[25] European Medicines Agency. *EMEA/H/C/000278 Herceptin: EPAR - Product Information*. URL: https://www.ema.europa.eu/en/documents/product-information/herceptin-epar-product-information_en.pdf (visited on 12/28/2022).

[26] European Medicines Agency. *EMEA/H/C/000406 - Glivec: EPAR - Product Information*. URL: https://www.ema.europa.eu/en/documents/product-information/glivec-epar-product-information_en.pdf (visited on 12/14/2022).

[27] European Medicines Agency. *EMEA/H/C/001016 - Iressa: EPAR - Product Information*. URL: https://www.ema.europa.eu/en/documents/assessment-report/iressa-epar-public-assessment-report_en.pdf (visited on 12/30/2022).

[28] European Medicines Agency. *EMEA/H/C/004124- Tagrisso: EPAR - Product Information*. URL: https://www.ema.europa.eu/en/documents/product-information/tagrisso-epar-product-information_en.pdf (visited on 12/14/2022).

[29] European Medicines Agency Web site. *Biomarker*. URL: https://www.ema.europa.eu/en/glossary/biomarker (visited on 12/14/2022).

[30] FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and other Tools)*. 2016-. URL: https://www.ncbi.nlm.nih.gov/books/NBK326791/ (visited on 12/14/2022).

[31] Fehring, R. J., Schneider, M., Raviele, K., Rodriguez, D., and Pruszynski, J. "Randomized comparison of two Internet-supported fertility-awareness-based methods of family planning". In: *Contraception* 88.1 (2013), pp. 24–30.

[32] Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., and Kelderman, H. "Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees". In: *Behavior Research Methods* 50.5 (2018), pp. 2016–2034.

[33] Food and Drug Administration. *Enrichment strategies for clinical trials to support approval of human drugs and biological products*. 2019. URL: https://www.fda.gov/media/121320/download (visited on 12/15/2022).

[34] Food and Drug Administration. *NDA 208065: Osimertinib (TAGRISSO): Labeling-Package Insert*. URL: https://www.accessdata.fda.gov/drugsatfda_docs/label/2022/208065s027lbl.pdf (visited on 12/14/2022).

[35] Food and Drug Administration Web site. *Precision medicine*. URL: https://www.fda.gov/medical-devices/in-vitro-diagnostics/precision-medicine (visited on 12/15/2022).

[36] Friede, T., Parsons, N., and Stallard, N. "A conditional error function approach for subgroup selection in adaptive clinical trials". In: *Statistics in Medicine* 31.30 (2012), pp. 4309–4320.

[37] Friede, T., Stallard, N., and Parsons, N. "Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R". In: *Biometrical Journal* 62.5 (2020), pp. 1264–1283.

[38] Hansen, B. E. "Approximate Asymptotic P Values for Structural-Change Tests". In: *Journal of Business & Economic Statistics* 15.1 (1997), pp. 60–67.

[39] Heyard, R., Timsit, J.-F., Essaied, W. I., and Held, L. "Dynamic clinical prediction models for discrete time-to-event data with competing risks-A case study on the OUTCOMEREA database". In: *Biometrical Journal* 61.3 (2019), pp. 514–534.

[40] Hjort, N. L. and Koning, A. "Tests For Constancy Of Model Parameters Over Time". In: *Journal of Nonparametric Statistics* 14.1-2 (2002), pp. 113–132.

[41] Hothorn, T. and Zeileis, A. "partykit: A Modular Toolkit for Recursive Partytioning in R". In: *Journal of Machine Learning Research* 16.118 (2015), pp. 3905–3909.

[42] Huber, C., Benda, N., and Friede, T. "A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations". In: *Pharmaceutical Statistics* 18.5 (2019), pp. 600–626.

[43] Huber, C., Benda, N., and Friede, T. "Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning". In: *Advances in Data Analysis and Classification* 16.3 (2022), pp. 797–815.

[44] Huber, C., Friede, T., Stingl, J., and Benda, N. "Classification of Companion Diagnostics: A New Framework for Biomarker-Driven Patient Selection". In: *Therapeutic Innovation & Regulatory Science* 56.2 (2022), pp. 244–254.

[45] Huber, C., Schmid, M., and Friede, T. *Model-based recursive partitioning for discrete event times*. (2022). URL: http://arxiv.org/pdf/2209.06592v1.

[46] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials*. 2020. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf (visited on 12/14/2022).

[47] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). *Statistical Principles for Clinical Trials E9*. 1998. URL: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf (visited on 12/14/2022).

[48] Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., and White, I. R. "A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio". In: *Statistics in Medicine* 37.7 (2018), pp. 1059–1085.

[49] Jenkins, M., Stone, A., and Jennison, C. "An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints". In: *Pharmaceutical Statistics* 10.4 (2011), pp. 347–356.

[50] Johnston, S. E., Lipkovich, I., Dmitrienko, A., and Zhao, Y. D. "A two-stage adaptive clinical trial design with data-driven subgroup identification at interim analysis". In: *Pharmaceutical Statistics* 21.5 (2022), pp. 1090–1108.

[51] Kontopantelis, E. "A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study". In: *Research Synthesis Methods* 9.3 (2018), pp. 417–430.

[52] Kontopantelis, E. "A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study". In: *Research Synthesis Methods* 9.3 (2018), pp. 417–430.

[53] Langer, C. J. "Roles of EGFR and KRAS Mutations in the Treatment Of Patients With Non-Small-Cell Lung Cancer". In: *P & T : a peer-reviewed journal for formulary management* 36.5 (2011), pp. 263–279.

[54] LeBlanc, M., Moon, J., and Crowley, J. "Adaptive Risk Group Refinement". In: *Biometrics* 61.2 (2005), pp. 370–378.

[55] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3 (2016).

[56] Lipkovich, I. and Dmitrienko, A. "Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES". In: *Journal of Biopharmaceutical Statistics* 24.1 (2014), pp. 130–153.

[57] Lipkovich, I., Dmitrienko, A., and B. D'Agostino Sr., R. "Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials". In: *Statistics in Medicine* 36.1 (2017), pp. 136–196.

[58] Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. "Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations". In: *Statistics in Medicine* 30.21 (2011), pp. 2601–2621.

[59] Loh, W.-Y., Cao, L., and Zhou, P. "Subgroup identification for precision medicine: A comparative review of 13 methods". In: *WIREs Data Mining and Knowledge Discovery* 9.5 (2019).

[60] Longinetti, E. and Fang, F. "Epidemiology of amyotrophic lateral sclerosis: an update of recent literature". In: *Current opinion in neurology* 32.5 (2019), pp. 771–776.

[61] Mandrekar, S. J. and Sargent, D. J. "Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges". In: *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 27.24 (2009), pp. 4027–4034.

[62] Mistry, D., Stallard, N., and Underwood, M. "A recursive partitioning approach for subgroup identification in individual patient data meta-analysis". In: *Statistics in Medicine* 37.9 (2018), pp. 1550–1561.

[63] Moradian, H., Yao, W., Larocque, D., Simonoff, J. S., and Frydman, H. "Dynamic estimation with random forests for discrete–time survival data". In: *Canadian Journal of Statistics* 50.2 (2022), pp. 533–548.

[64] Neuhaus, J. M., McCulloch, C. E., and Boylan, R. "Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes". In: *Statistics in medicine* 32.14 (2013), pp. 2419–2429.

[65] Neyman, J and Scott, E. "Consistent Estimates Based on Partially Consistent Observations". In: *Econometrica* 16.1 (1948), p. 1.

[66] Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., and Posch, M. "Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review". In: *Journal of Biopharmaceutical Statistics* 26.1 (2016), pp. 99–119.

[67] Patel, S, Hee, S., Mistry, D, Jordan, J, Brown, S, Dritsaki, M, Ellard, D, Friede, T., Lamb, S., Lord, J, et al. "Identifying back pain subgroups; developing and applying approaches using individual patient data collected within clinical trials". In: *Programme Grants for Applied Research* 4.10 (2016), pp. 1–314.

[68] Patel, S., Hee, S. W., Mistry, D., Jordan, J., Brown, S., Dritsaki, M., Ellard, D, Friede, T., Lamb, S. E., Lord, J., et al. "Identifying back pain subgroups; developing and applying approaches using individual patient data collected within clinical trials". In: *Programme Grants for Applied Research* 4.10 (2016), pp. 1–314.

[69] Personalized Medicine Coalition. *Personalized medicine at FDA The Scope an Significance of Progress in 2021*. 2021. URL: `https://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/Personalized_Medicine_at_FDA_The_Scope_Significance_of_Progress_in_2021.pdf` (visited on 12/14/2022).

[70] Phillips, R., Sauzet, O., and Cornelius, V. "Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy". In: *BMC medical research methodology* 20.1 (2020), p. 288.

[71] Placzek, M. and Friede, T. "Blinded sample size recalculation in adaptive enrichment designs". In: *Biometrical Journal* (2022).

[72] Riley, R. D., Legha, A., Jackson, D., Morris, T. P., Ensor, J., Snell, K. I. E., White, I. R., and Burke, D. L. "One-stage individual participant data meta-analysis models for continuous and binary outcomes: Comparison of treatment coding options and estimation methods". In: *Statistics in Medicine* 39.19 (2020), pp. 2536–2555.

[73] Scheike, T. H. and Keiding, N. "Design and analysis of time-to-pregnancy". In: *Statistical Methods in Medical Research* 15.2 (2006), pp. 127–140.

[74] Schleidgen, S., Klingler, C., Bertram, T., Rogowski, W. H., and Marckmann, G. "What is personalized medicine: sharpening a vague term based on a systematic literature review". In: *BMC Medical Ethics* 14 (2013), p. 55.

[75] Schmid, M. and Berger, M. "Competing risks analysis for discrete time–to–event data". In: *WIREs Computational Statistics* 13.5 (2021).

[76] Schmid, M., Küchenhoff, H., Hoerauf, A., and Tutz, G. "A survival tree method for the analysis of discrete event times in clinical and epidemiological studies". In: *Statistics in Medicine* 35.5 (2016), pp. 734–751.

[77] Seibold, H., Zeileis, A., and Hothorn, T. "Model-Based Recursive Partitioning for Subgroup Analyses". In: *The International Journal of Biostatistics* 12.1 (2016), pp. 45–63.

[78] Sela, R. J. and Simonoff, J. S. "RE-EM trees: a data mining approach for longitudinal and clustered data". In: *Machine Learning* 86.2 (2012), pp. 169–207.

[79] Sies, A. and Van Mechelen, I. "Comparing Four Methods for Estimating Tree-Based Treatment Regimes". In: *The International Journal of Biostatistics* 13 (2017).

[80] Simmonds, M., Stewart, G., and Stewart, L. "A decade of individual participant data meta-analyses: A review of current practice". In: *Contemporary Clinical Trials* 45.Pt A (2015), pp. 76–83.

[81] Steinberg, J. S., Göbel, A. P., Thiele, S., Fleckenstein, M., Holz, F. G., and Schmitz-Valckenberg, S. "Development of intraretinal cystoid lesions in eyes with intermediate age-related macular degeneration". In: *Retina* 36.8 (2016), pp. 1548–1556.

[82] Strobl, C. and Leisch, F. "Against the "one method fits all data sets" philosophy for comparison studies in methodological research". In: *Biometrical Journal* (2022).

[83] Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. "Subgroup Analysis via Recursive Partitioning". In: *Journal of Machine Learning Research* 10.5 (2009), pp. 141–158.

[84] Sun, S., Sechidis, K., Chen, Y., Lu, J., Ma, C., Mirshani, A., Ohlssen, D., Vandemeulebroecke, M., and Bornkamp, B. "Comparing algorithms for characterizing treatment effect heterogeneity in randomized trials". In: *Biometrical Journal* (2022).

[85] Tanniou, J., van der Tweel, I., Teerenstra, S., and Roes, K. C. "Level of evidence for promising subgroup findings in an overall non-significant trial". In: *Statistical Methods in Medical Research* 25.5 (2016), pp. 2193–2213.

[86] The International Weight Management in Pregnancy (i-WIP) Collaborative Group. "Effect of diet and physical activity based interventions in pregnancy on gestational weight gain and pregnancy outcomes: meta-analysis of individual participant data from randomised trials". In: *BMJ (Clinical research ed.)* 358 (2017), j3119.

[87] *The Precision Medicine Initiative*. URL: `https://obamawhitehouse.archives.gov/precision-medicine` (visited on 12/30/2022).

[88] Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), 267–88.

[89] Tutz, G. and Schmid, M. *Modeling Discrete Time-to-Event Data*. Springer Series in Statistics. Cham: Springer, 2016.

[90] Wang, S.-J. "Utility of adaptive strategy and adaptive design for biomarker-facilitated patient selection in pharmacogenomic or pharmacogenetic clinical development program". In: *Journal of the Formosan Medical Association = Taiwan yi zhi* 107.12 Suppl (2008), pp. 19–27.

[91] Wang, X. V., Cole, B., Bonetti, M., and Gelber, R. D. "Meta-STEPP: subpopulation treatment effect pattern plot for individual patient data meta-analysis". In: *Statistics in Medicine* 35.21 (2016), pp. 3704–3716.

[92] Wang, X. V., Cole, B., Bonetti, M., and Gelber, R. D. "Meta-STEPP with random effects". In: *Research Synthesis Methods* 9.2 (2018), pp. 312–317.

[93] Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., and Klein, T. E. "An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine". In: *Clinical Pharmacology and Therapeutics* 110.3 (2021), pp. 563–572.

[94] Willett, J. B. and Singer, J. D. "Investigating onset, cessation, relapse, and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurrence". In: *Journal of Consulting and Clinical Psychology* 61.6 (1993), pp. 952–965.

[95] Zeileis, A. and Hornik, K. "Generalized M-fluctuation tests for parameter instability". In: *Statistica Neerlandica* 61.4 (2007), pp. 488–508.

[96] Zeileis, A., Hothorn, T., and Hornik, K. "Model-Based Recursive Partitioning". In: *Journal of Computational and Graphical Statistics* 17.2 (2008), pp. 492–514.

Appendix A

Original Articles

A.1 Classification of Companion Diagnostics: A new framework for biomarker-driven patient selection

The published version is publicly available from
https://doi.org/10.1007/s43441-021-00352-2.

A.2 A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations

The published version is publicly available from
https://doi.org/10.1002/pst.1951.

A.3 Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning

The published version is publicly available from
https://doi.org/10.1007/s11634-021-00458-3.

A.4 Model-based recursive partitioning for discrete event times

The preprint is publicly available from http://arxiv.org/pdf/2209.06592v1.